DRDC No. CR 2003-060

DEVELOPMENT OF GENERIC AIRCREW MEASURES OF PERFORMANCE FOR DISTRIBUTED MISSION TRAINING

by:

Michael L. Matthews, Tabbeus M. Lamoureux

Human*systems*, Incorporated
111 Farquhar St., 2nd floor
Guelph, ON N1H 3N4

Project Manager:
Kim Iwasa-Madge
(519) 836 5911

PWGSC Contract No. W7711-007694/001/TOR
Call-Up 7694-06

On behalf of
DEPARTMENT OF NATIONAL DEFENCE

as represented by
Defence Research and Development Canada
1133 Sheppard Avenue West
Toronto, Ontario, Canada
M3M 3B9

DRDC-Toronto Scientific Authority
Stuart Grant
(416) 635-2000

31 March 2003

104-05538/

© HER MAJESTY THE QUEEN IN RIGHT OF CANADA (2003)
as represented by the Minister of National Defense

© SA MAJESTE LA REINE EN DROIT DUE CANADA (2003)
Defense Nationale Canada

Abstract

Advances in technology have made simulation and, latterly, distributed mission simulation valuable additions to the training of aircrew. Simulation is widely accepted by the aviation community and much research exists to show the benefits and most profitable applications of simulation. Distributed mission training represents an enhancement to simulation although at this point it is unproven exactly what training objectives should be associated with it and what additional benefits will accrue when compared to traditional simulation or flying training. This project developed generic measures of performance for application to distributed mission training exercises. The application of these measures of performance will allow training organisations to make valid statements about the benefits of distributed mission training and informed decisions to be made regarding which training objectives to address through the use of distributed mission training.

Human*systems*® Incorporated were tasked with reviewing literature provided by DRDC Toronto in order to identify potential measures of performance. In particular, the Scientific Authority was interested in measures of mission planning, mission execution, mission debriefing, situation awareness and the change in aircrew knowledge structures, as relevant to distributed mission training. A measurement model has been developed that includes a conceptual outline of a CF-18 mission, a behavioural hierarchy composed of domains, categories and specific behaviours, and a range of associated rating scales and objective measures. Additionally, a trial plan for the application of one particular measure (Pathfinder – description and measurement of knowledge structures) has been developed.

Résumé

Grâce aux avancées technologiques, la simulation et, par la suite, la simulation de mission à distance, sont devenues de précieux ajouts à l'instruction de l'équipage. La simulation est largement répandue dans le monde de l'aviation, et bien des recherches démontrent les avantages et les applications les plus profitables de la simulation. L'instruction de mission à distance représente une amélioration en matière de simulation, même si l'on ne connaît pas encore exactement les objectifs d'instruction qu'il faudrait y associer et les avantages supplémentaires qui en découleront par rapport à la simulation traditionnelle ou à l'entraînement au vol. Les responsables de ce projet ont élaboré des mesures du rendement génériques applicables aux exercices d'instruction de mission à distance. L'application de ces mesures du rendement permettra aux organismes de formation de faire des recommandations pertinentes sur les avantages de l'instruction de mission à distance et de prendre des décisions éclairées sur les objectifs d'instruction à viser dans le cadre de l'instruction de mission à distance.

Human*systems*® Incorporated a été chargée d'examiner la documentation fournie par RDDC Toronto en vue de déterminer des mesures du rendement potentielles. En particulier, le responsable des questions scientifiques s'intéressait aux mesures relatives à la planification de mission, l'exécution de mission, le debriefing de mission, la connaissance de la situation, ainsi que les modifications au niveau des structures de connaissances de l'équipage, en ce qui concerne l'instruction de mission à distance. On a élaboré un modèle de mesure incluant un aperçu conceptuel d'une mission de CF-18, une hiérarchie du comportement composée de domaines, de catégories et de comportements particuliers, ainsi qu'une gamme d'échelles de notation et de mesures objectives connexes. On a également élaboré un plan d'essai pour l'application d'une mesure particulière (Pathfinder – description et mesure des structures de connaissances).

Executive Summary

DRDC Toronto is becoming involved in several large-scale international Distributed Mission Training (DMT) exercises. Such exercises, while less costly than the use of aircraft assets, still require a significant effort to organise and run, especially in the international context. As a result, these events are likely to be relatively few in number and provide significant challenges to collecting useful human performance data. Additionally, no two DMT trials will be exactly alike. Under the auspices of the Advanced Distributed Mission Training (ADMT) Technology Demonstration (TD) project, DRDC Toronto is investigating Measures of Performance (MOPs) that can be compared across dissimilar DMT events. As a starting point, three elements have been identified that are common among all air missions: planning and briefing; situational awareness during the mission; and post-mission debriefing. This contract investigates the critical behaviours and functions involved in these elements and developing practical MOPs that can be applied across dissimilar DMT events to gain insight into the training value of DMT, or for other assessment purposes.

DRDC Toronto have also indicated an interest in the utility of the Pathfinder technique for describing and assessing knowledge structures and the changes that they may undergo during training. Thus, an additional goal of the contract is to investigate the advantages and disadvantages associated with use and application of the Pathfinder technique in the context of DMT and generic Air Force (AF) training.

In order to develop MOPs and investigate the Pathfinder technique, Human*systems* Incorporated® undertook a number of tasks:

- Literature review of 95 papers on the subject of MOPs, Pathfinder, Situation Awareness, Mission Essential Competencies, planning, debriefing and Crew Resource Management;
- Familiarisation with DMT and AF mission sequences through visits to the US Air Force Research Laboratory (USAFRL) and Defence Research and Development (DRDC) Toronto;
- Workshop on Pathfinder with the creator of the technique;
- Review of previous Human*systems* Incorporated® work on MOPs.

In total, these tasks provided large amounts of information that were considered in the Canadian DMT context and distilled to a measurement framework. However, it should be noted that USAFRL efforts in this arena have used task analytic techniques to decompose mission elements to a point where diagnostic MOPs, sensitive to changes in expertise and mission demands, could be identified. This contract effectively considered fragments of these task analyses to generate generic MOPs for the Canadian AF and used Canadian subject matter experts (SMEs) to refine the task selection.

The measurement framework stems from a conceptual model of an AF mission. This model has 9 stages but is further grouped into three broad mission categories for the purposes of defining MOPs. These stages are Mission Planning, Mission Execution and Mission Debriefing. The Mission Execution phases are also grouped into three categories for convenience. These categories are Start/Taxi/Takeoff (STTO); Transit to Tactical Rendezvous Point (TRP); and Ingress/Execute/Egress (IEXEG). This mission model is suitably generic to be applied to any AF mission.

For each broad mission category, behaviour domains of interest are identified. Corresponding behaviour categories for each behaviour domain are then identified, leading to the generation of specific behaviours of interest. These specific behaviours of interest are the items that need to be

measured during DMT exercises. Again, these specific behaviours are generic so they may be applied to any AF mission.

For measurement purposes, three measurement approaches have been recommended: Objective performance data; Behaviourally Anchored Rating Scales (BARS) scored by SMEs; and the Pathfinder technique. This recommendation is based upon credible evidence which shows that BARS and objective performance data, applied correctly, provide valid, reliable and diagnostic output. The selected measures are applied to mission planning, situation awareness during the mission, and mission debriefing

It is recommended that the Pathfinder technique be applied to the assessment of the development and change of knowledge structures. The validity and reliability of Pathfinder has not been systematically demonstrated, and it has not been compared with other similar techniques, although the face validity of Pathfinder solutions has been demonstrated in a number of studies. The relationship of Pathfinder-derived knowledge structures to actual task performance still needs to be established, as does the test-retest reliability of the various statistical outputs available.

A complete measurement model is presented, based on the above activities and findings. The complete list of BARS and objective performance data requirements recommended by this project are provided in Annex A of this report. Further, a trial plan is presented in Annex B of this report for using Pathfinder to describe knowledge structures before and after training.

This report concludes that further work should be undertaken to test the recommended MOPs during actual simulation exercises and, if possible, DMT exercises. Further, this report concludes that to develop additional effective MOPs, task analysis needs to be done on a range of mission types in order to decompose mission training to a point where sensitive and diagnostic MOPs can be specified.

Sommaire

RDDC Toronto participe à plusieurs exercices internationaux d'instruction de mission à distance (IMD) de grande envergure. Ces exercices sont moins dispendieux que l'utilisation d'aéronefs, mais leur organisation et leur déroulement exigent un apport important, surtout dans le contexte international. Ces activités sont donc appelées à être relativement peu nombreuses, et la collecte de données utiles sur les performances humaines représente des défis considérables. De plus, on ne trouvera pas deux essais d'IMD exactement semblables. Dans le cadre du Projet de démonstration de technologies (DT) de l'instruction de mission avancée à distance (IMAD), RDDC Toronto étudie des mesures du rendement (MR) que l'on peut comparer entre des activités d'IMD différentes. On a repéré au départ trois éléments communs à toutes les missions aériennes : la planification et l'exposé de mission; la connaissance de la situation pendant la mission; et le debriefing après une mission. Ce marché a pour but d'examiner les fonctions et les comportements cruciaux rattachés à ces éléments et d'élaborer des MR pratiques que l'on peut appliquer à des activités d'IMD différentes pour avoir un aperçu de la valeur de l'IMD du point de vue de l'instruction, ou pour d'autres évaluations.

RDDC Toronto a également manifesté un intérêt envers l'utilité de la technique Pathfinder pour décrire et évaluer les structures de connaissances et les changements susceptibles de se produire pendant l'instruction. Le marché a donc aussi pour but d'examiner les avantages et les inconvénients liés à l'utilisation et à l'application de la technique Pathfinder dans le contexte de l'IMD et de l'instruction générique de la Force aérienne (FA).

Humansystems IncorporatedMD a entrepris diverses tâches, dans le but d'élaborer des MR et d'examiner la technique Pathfinder :

- Analyse documentaire de 95 documents portant sur les MR, Pathfinder, la connaissance de la situation, les compétences essentielles à la mission, la planification, le debriefing et la gestion du personnel affecté aux aéronefs
- Familiarisation avec l'IMD et les séquences de mission de la FA, en se rendant à l'US Air Force Research Laboratory (USAFRL) et à Recherche et développement pour la défense Canada (RDDC), à Toronto
- Atelier sur Pathfinder avec le créateur de la technique
- Examen des travaux précédents de Human*systems*® Incorporated sur les MR.

En tout, ces tâches ont permis de recueillir d'importantes sommes de renseignements, que l'on a examinés dans le contexte canadien de l'IMD et intégrés à un cadre de mesure. Il est toutefois à remarquer que dans ce domaine, les responsables de l'USAFRL ont employé des techniques d'analyse des tâches pour décomposer les éléments de mission de manière à pouvoir identifier des MR de diagnostic adaptés aux changements au niveau de l'expertise et des exigences des missions. Dans le cadre de ce marché, on a examiné efficacement des fragments de ces analyses de tâche en vue de produire des MR génériques pour l'Aviation canadienne, et fait appel à des spécialistes canadiens en la matière pour préciser la sélection des tâches.

Le cadre de mesure découle d'un modèle conceptuel de mission de la FA. Ce modèle englobe neuf étapes, qui se regroupent en trois grandes catégories de missions permettant de définir des MR. Ces étapes sont la planification de mission, l'exécution de mission et le debriefing de mission. Les phases d'exécution de mission se regroupent également en trois catégories pour des raisons de commodité. Ces catégories sont lancement/roulage/décollage (STTO); passage à un point de rendez-vous tactique

(TRP); et entrée/exécution/sortie (IEXEG). Ce modèle de mission est générique et peut s'appliquer à n'importe quelle mission de la FA.

Pour chaque grande catégorie de mission, on détermine des domaines d'intérêt en matière de comportement. On établit ensuite des catégories de comportement correspondant à chaque domaine de comportement, afin de dégager des comportements particuliers d'intérêt. Ces comportements particuliers d'intérêt sont les éléments à mesurer au cours des exercices d'IMD. Là encore, ces comportements particuliers sont génériques, de manière à pouvoir s'appliquer à n'importe quelle mission de la FA.

On a recommandé trois démarches en vue d'effectuer des mesures : les données de rendement objectif; les échelles d'évaluation fondées sur le comportement (EEFC) notées par des experts en la matière; et la technique Pathfinder. Cette recommandation est basée sur une preuve crédible qui démontre que l'EEFC et les données sur le rendement objectif, lorsqu'elles sont appliquées correctement, fournissent un produit de diagnostic pertinent et fiable. Les mesures choisies s'appliquent à la planification de mission, à la connaissance de la situation pendant la mission, et au debriefing sur la mission.

Il est conseillé d'appliquer la technique Pathfinder à l'évaluation de l'élaboration et de la modification des structures des connaissances. La validité et la fiabilité de Pathfinder n'ont pas été démontrées systématiquement, et on ne l'a pas comparée à d'autres techniques semblables, mais beaucoup d'études ont prouvé la validité apparente des solutions Pathfinder. Il reste à établir les liens entre les structures des connaissances qui découlent de Pathfinder et l'exécution réelle des tâches, ainsi que la fiabilité de test-retest des divers produits statistiques disponibles.

Nous présentons un modèle de mesure complet, basé sur les activités et les conclusions exposées ci-dessus. La liste complète des EEFC et des exigences en matière de données sur le rendement objectif recommandées dans le cadre de ce projet figure à l'annexe A de ce rapport. De plus, un plan d'essai est présenté à l'annexe B de ce rapport en vue d'employer Pathfinder pour décrire les structures de connaissances avant et après l'instruction.

Nous concluons dans ce rapport qu'il faudrait entreprendre d'autres études pour mettre à l'essai les MR recommandées au cours d'exercices de simulation réels et, si possible, d'exercices d'IMD. Nous concluons également dans ce rapport que pour élaborer d'autres MR efficaces, il faudrait procéder à une analyse des tâches relatives à une gamme de types de missions, en vue de décomposer l'instruction de mission et de pouvoir ainsi préciser des MR sensibles et de diagnostic.

Table of Contents

TABLE OF CONTENTS I

ABSTRACT I

RÉSUMÉ II

EXECUTIVE SUMMARY III

SOMMAIRE V

1. BACKGROUND 1

1.1 CONTRACT NUMBER 1
1.2 OBJECTIVES 1
1.3 SCIENTIFIC AUTHORITY 2
1.4 LIST OF ACRONYMS 2
1.5 ACKNOWLEDGMENTS 3

2. STRUCTURE OF REPORT 5

3. PROCESS USED 7

3.1 LITERATURE REVIEW 7
3.2 DMT FAMILIARIZATION 7
3.3 MISSION FAMILIARIZATION 7
3.4 PATHFINDER WORKSHOP 7
3.5 DISCUSSIONS WITH USAFRL SCIENTISTS 8
3.6 ADOPTION AND ADAPTATION OF SUITABLE MOPS FROM LITERATURE 8

4. THE MEASUREMENT MODEL 9

4.1 MODEL FRAMEWORK 9
4.1.1 Why measure *9*
4.1.2 Selection of what to measure *9*
4.1.3 Identifying the behaviour domains/constructs *9*
4.1.4 Defining specific behaviours *9*
4.1.5 Selection of measures and procedures *10*
4.1.6 Defining measurement Conditions *10*
4.1.7 Collect data *10*
4.1.8 Analyse data *10*

5. MEASUREMENT FRAMEWORK 13

5.1 REASONS FOR MEASUREMENT IN PRESENT CONTEXT 13
5.2 THE DMT DOMAIN 13
5.3 PLANNING, MISSION EXECUTION AND DEBRIEFING 14
5.4 THE BEHAVIOUR DOMAIN 15
5.4.1 Mission planning *15*
5.4.2 Mission execution *16*
5.4.3 Mission debriefing *16*
5.5 THE SPECIFIC CONSTRUCTS 16

5.5.1 Mission planning *16*
5.5.2 Mission execution *18*
5.5.3 Mission debriefing *19*
5.6 THE SPECIFIC BEHAVIOURS TO MEASURE 20
5.6.1 Mission planning *20*
5.6.2 Mission execution *24*
5.6.3 Mission debriefing *25*

6. MEASURES OF PERFORMANCE 27

6.1 SUMMARY LITERATURE REVIEW 27
6.2 RECOMMENDATIONS FOR MEASURES 28
6.3 OBJECTIVE PERFORMANCE DATA 28
6.4 BEHAVIOURALLY ANCHORED RATING SCALES (BARS) 29
6.4.1 Issues for the administration of BARS *31*
6.5 PATHFINDER 32
6.5.1 Overview and historical development *32*
6.5.2 Details of core concepts *32*
6.5.3 Steps in the Pathfinder approach *33*
6.5.4 Examples of use in training and other similar environments *35*
6.5.5 Limitations and issues not fully resolved *36*
6.5.6 Application to Distributed Mission Training *39*
6.6 SITUATION AWARENESS: CONCEPTS AND MEASUREMENT 39
6.6.1 Definitions *39*
6.6.2 The work of Endsley *41*
6.6.3 Suggested operational definition of SA *42*
6.6.4 The Measurement of SA *44*
6.6.5 Subjective techniques: I. Post-event retrospective recall *57*
6.6.6 Subjective techniques: II. During task execution *57*
6.6.7 Subjective techniques: III. SME ratings of behaviour *59*

7. THE COMPLETE MEASUREMENT MODEL 67

8. MEASUREMENT PROCEDURES AND LOGISTICS 71

8.1 DATA CAPTURE 71
8.2 PERSONNEL 72
8.3 WHEN TO MEASURE? 72
8.4 WHO TO MEASURE? 72

9. MEASUREMENT CONDITIONS 75

9.1 DATA COLLECTION 75
9.2 DATA ANALYSIS 75

10. CONCLUSIONS 77

11. REFERENCES AND SOURCES CONSULTED 79

ANNEX A: RECOMMENDED MEASURES OF PERFORMANCE (MOPS) 1

BEHAVIOURALLY ANCHORED RATINGS SCALES (BARS) 5
Mission Planning and preparation *5*
Mission Execution *16*
MISSION DEBRIEF 32
OBJECTIVE PERFORMANCE DATA 37
Mission Planning *37*

Mission Execution *38*

ANNEX B: PROPOSED METHODOLOGY/TRIAL PLAN FOR USE OF PATHFINDER IN EVALUATING AIRCREW TRAINING **1**

1. DOMAIN SELECTION 1
2. CONCEPT SELECTION 1
3. CONCEPT REFINEMENT 2
4. PRE-TRAINING ADMINISTRATION OF PAIRWISE PROCEDURE 2
5. POST-TRAINING ADMINISTRATION OF PAIRWISE PROCEDURE 3
6. STATISTICAL ANALYSIS AND INTERPRETATION OF DATA 3
7. REPORTING 3
8. DETAILED OUTLINE OF RESOURCES REQUIRED 4

List of Tables and Figures

Table 1: Mapping of mission planning behaviour categories to behaviour domains 18

Table 2: Mapping of mission execution behaviour categories to behaviour domains 19

Table 3: Mapping of mission debriefing behaviour categories to behaviour domains 20

Table 4: Core Principles of Planning 21

Table 5: Examples of Effective Mission Preparation Behaviours (from Spiker, Nullmeyer and Tourville, 2000) 22

Table 6: Mapping of mission planning specific behaviours to behaviour categories 23

Table 7: Mapping of mission execution specific behaviours to behaviour categories 25

Table 8: Mapping of mission debriefing specific behaviours to behaviour categories 26

Table 9: Summary evaluation of situation awareness measures 62

Table 10: Spatial, temporal and identity awareness 64

Table 11: System awareness 64

Table 12: Resource and crew awareness 65

Table 13: Tactical awareness during engagement 65

Table 14: Mission/goal awareness 66

Table 15: Measurement Concept: Planning and Briefing 67

Table16: Measurement Concept: Mission Execution 68

Table 17: Measurement Concept: Debriefing 69

Table 18: Measurement Points and Simulation Participants 73

Figure 1: Conceptual Mission Model 14

Figure 2: Example BARS 30

Figure 3: Situation Awareness process and Measurement Model (adapted from Endsley (1996)) 45

Figure 4: Example mission execution BARS 60

1. Background

DRDC Toronto is becoming involved in several large-scale international Distributed Mission Training (DMT) exercises. Such exercises, while less costly than the use of aircraft assets, still require a significant effort to organise and run, especially in the international context. As a result, these events are likely to be relatively few in number and provide significant challenges to collecting useful human performance data. Additionally, no two DMT trials will be exactly alike. Under the auspices of the Advanced Distributed Mission Training (ADMT) Technology Demonstration (TD) project, DRDC Toronto is investigating Measures of Performance (MOPs) that can be compared across dissimilar DMT events. As a starting point, three elements have been identified that are common among all air missions: planning and briefing; situational awareness during the mission; and post-mission debriefing. This contract will investigate the critical behaviours and functions involved in these elements and developing practical MOPs that can be applied across dissimilar DMT events to gain insight into the training value of DMT, or for other assessment purposes.

DRDC Toronto have also indicated an interest in the utility of the Pathfinder technique for describing and assessing knowledge structures and the changes that they may undergo during training. Thus, an additional goal of the contract is to investigate the advantages and disadvantages associated with use and application of the Pathfinder technique in the context of DMT and generic Air Force (AF) training.

There are two distinct, potential audiences for this report - defence scientists and the AF community. Their needs in terms of detail of information provided and the purposes to which MOPs will be put are somewhat different. We have attempted to address these different needs by making the main body of the text address directly the needs of Defence Scientists by providing rationale, explanations, references to literature and detailed analyses. For the AF community, we have provided an Annex of MOPs for specific military functions with some guidance as to how they should be implemented and without much explanation for their derivation.

1.1 Contract Number

This work was commissioned by DRDC Toronto with Human*systems* Incorporated® through standing offer W7711-007694/001/TOR.

1.2 Objectives

The objective of this work was to review provided literature on aircrew MOPs to evaluate DMT and to recommend generic MOPs and methods applicable across different mission and training contexts.

Specifically, the tasking is to make recommendations should be made in the following areas:

- MOPs for mission planning;
- MOPs for mission debriefing;
- MOPs for situation awareness during a mission;
- Application of the Pathfinder technique.

One further objective was to develop a trial plan for the application of the Pathfinder technique to the assessment of crew[1] knowledge structures in training.

1.3 Scientific Authority

The Scientific Authority for this contract was initially Ian Mack, DRDC Toronto. Due to the Scientific Authority taking up a new role, the Scientific Authority from the beginning of March, 2003 onward was Stuart Grant, DRDC Toronto.

1.4 List of Acronyms

ADMT	Advanced Distributed Mission Training
AF	Air Force
AFTL	Air Force Task List
BARS	Behaviourally-Anchored Rating Scales
CMTR	Coalition Mission Training Research
CRM	Crew Resource Management
DMT	Distributed Mission Training
DRDC	Defence Research and Development Canada
EP	Embedded Probe
IEXEG	Ingress/Execute/Egress
JKI	Job Knowledge Inventory
KSEs	Knowledge, Skills and Experiences
MDS	Multidimensional Scaling
MECs	Mission Essential Competencies
MOEs	Measures Of Effectiveness
MOPs	Measures Of Performance
PEMDA	Post-Event Mission Data Analysis
PFNET	Pathfinder Net
RTB	Return To Base
SA	Situation Awareness
SAAS	Situation Assessment and Action Selection
SAGAT	Situation Awareness Global Assessment Technique
SJI	Situational Judgement Inventory
SMEs	Subject Matter Experts
STTO	Start, Taxi, Take-off
SWORD	Subjective Workload Dominance
TD	Technology Demonstration
TOLD	Take-Off and Landing Data
TRP	Transit to Tactical Rendezvous Point
TTWODLDE	Terrain, Target, Weather, Ordnance, Delivery, Lateral Support, Defences, Egress
USAF	** *defined as part of USAFRL not separately*
USAFRL	US Air Force Research Laboratory

[1] We use the word "crew" generically in this report to refer to any relevant mission-participating personnel. Thus, a crew could be a member of a team within a particular aircraft, or pilots of different aircraft that comprise a formation

1.5 Acknowledgments

We should like to acknowledge the important contributions to this report provided by Dr. Robert Nullmeyer of the USAFRL and Top Aces. Dr. Nullmeyer was generous in meeting our requests for specific examples of measures that were developed and proven in a USAF context, and which form the basis of the many of the measures provided in Annex A. Dr. Nullmeyer was also generous in meeting with us at length to discuss measurement issues, to give advice on practical matters and to generally discuss our own thoughts and ideas on how to extend and elaborate upon his own approaches.

We should also like to thank Paul Bouchard of Top Aces for their invaluable work in reviewing our measurement model for completeness, adapting it to a Canadian context and working with us to develop additional MOPs.

THIS PAGE INTENTIONALLY LEFT BLANK

2. Structure of Report

The report starts with an overview of the process used, then outlines in detail the specific measurement model adopted and the fills in the details of the measurement framework. The core contents of the report are found in the section on Measures of Performance, which contains a highly detailed review of the various measures with a special emphasis on Pathfinder and Situation Awareness.

The report then goes on to combine the MOPs into the measurement model and finishes with sections dealing procedures and measurement conditions.

The detailed MOPs themselves are to be found in Annex A, which may be read as a standalone document for users who just wish to know what measures to apply in a specific mission context. Thus, this annex could be used either by Defence Scientists or the Air Force training or operational analysis communities. For the rationale behind the measures, it will be necessary for the reader to consult the appropriate sections of text.

Annex B contains the detailed outline of a Trial Plan to implement the Pathfinder technology to evaluate a component of mission training. This outline contains the tasking and level of effort. The costs of the trial are provided separately in electronic format.

THIS PAGE INTENTIONALLY LEFT BLANK

3. Process used

3.1 Literature Review

A total of 95 items of documentation and literature were reviewed. Some of this documentation included internal USAF, USAFRL and Canadian Air Force documents, procedures, guidelines and manuals. The large majority of papers were published in the public domain in peer-reviewed journals and conference proceedings, notably the conference series 'Interservice-Industry Training, Simulation and Education Conference'.

The breakdown of papers according to the different measurement objectives is as follows:

- Mission Planning 8 papers
- Mission Performance 19 papers
- Mission Debriefing 1 paper
- Pathfinder 17 papers
- Situation Awareness 34 papers

Additionally, a number of further papers were reviewed due to their relevance and background material. These included 13 papers on the subject of Crew Resource Management (CRM) and 3 papers on the subject of 'mission essential competencies'.

3.2 DMT familiarization

Over the course of this contract, there were three opportunities to become familiar with DMT: during a trip to the USAFRL in Mesa, Arizona; during a joint US/UK/Canada DMT exercise; and during a briefing on the conduct and outcome of the joint exercise.

Familiarization focused on the facilities available to DMT participants, with special emphasis on planning tools, data recording facilities, and communications media between separate sites. Of additional interest was the manner in which these facilities are used by the participants.

3.3 Mission familiarization

Mission familiarization was gained largely through the literature. However, there were also opportunities to speak with pilots familiar with typical mission timetables and activities. This permitted the development of a conceptual model of a mission onto which measurement points and activities could be mapped.

3.4 Pathfinder workshop

During the two-day visit to the USAFRL in Mesa, Arizona a workshop was held covering the theory behind the Pathfinder technique, the development of suitable measurement concepts, the application of the technique and the interpretation of the data. The workshop was led by the developer of the technique, Roger Schvaneveldt, and involved the principle users of the technique at USAFRL. Since the USAFRL employ Pathfinder in the assessment of distributed mission training, their experience provided valuable insight for the application of Pathfinder to Canadian DMT.

3.5 Discussions with USAFRL Scientists

During the two day visit to the USAFRL the opportunity was taken to speak with their scientists. The personnel at USAFRL have a strong history of achievement in the field of simulation training, DMT, and the development of effective MOPs for aircrew.

Discussions with USAFRL scientists centred around the development of subjective and objective MOPs and the practical application of MOPs. A number of important issues were raised, including the so-called 'Friday Effect' where aircrew, after a week of training, are not as diligent as they should be in participating in subjective measurement sessions.

Discussions were also had around the future of simulation training and DMT. This provided valuable pointers to the future and resulted in the development of MOPs that will remain applicable for the foreseeable future.

3.6 Adoption and adaptation of suitable MOPs from literature

The development of suitable MOPs for Canadian CF-18 DMT was founded on training work in a number of different environments (e.g. simulated, live) and on a number of different platforms (e.g. MC-130, MH-53J, F-16). Consequently MOPs described in each paper had to be critically considered for their applicability to the CF-18 DMT environment and, further, to their more generic applicability to DMT simulations involving any air force weapons platform (as per the instructions of the Scientific Authority).

Most MOPs did ultimately come from the literature. However, several fundamental MOPs were uncovered during discussions with DMT researchers. Additionally, MOPs that had been distilled from the literature were made applicable to the Canadian Air Force and/or generic amongst weapons platform after discussion with researchers, SMEs and after visiting DMT facilities.

Because of the subjective and qualitative nature of most of the MOPs reported in the literature it was necessary to develop behavioural examples in order that the validity of the MOPs and the reliability of measurement be assured. Thus, MOPs developed in this report represent those reported in the literature, augmented through discussions with experts and in light of familiarisation with the facilities available. This approach helps to ensure that the MOPs are generic to air force DMT simulations in Canada, are reliable and valid, and provide added value to DMT exercises.

4. The Measurement Model

4.1 Model Framework

4.1.1 Why measure

Whatever the approach adopted, it is important that the purpose and goals of measurement drive the methodology. Thus, the initial questions must always be "why are we measuring" and "what are we measuring". The reasons for measuring are complex and varied but may include the following:

- The assessment of individuals against a performance standard
- Performance comparisons among teams
- Evaluation of different training approaches - technological and processes
- Evaluation of new systems
- Evaluation of interface designs
- Comparison of different operational procedures
- Scientific enquiry of underlying theoretical and conceptual issues

In general, we recommend the following structured approach to MOP identification and development that is recommended by Spiker and Nullmeyer (1995), which we have adapted for present purposes.

4.1.2 Selection of what to measure

The framework sets out the behavioural sub-fields of interest, and these in turn will specify the behaviour categories and constructs to be measured. In the present context, the overall framework involves processes of Planning, Mission Execution and Debriefing.

4.1.3 Identifying the behaviour domains/constructs

These are the primary functions and processes that comprise behaviour domains. Normally the identification of these would be achieved through mission, function and task analyses. Since such activities were beyond the scope of the current contract, and significant effort had already been expended in these areas by the USAFRL and others, we have relied on the literature reviewed and HSI experience in developing MOPs for other relevant military contexts. Examples of the behaviour domains within Mission Planning are: briefings, detailed planning, development of planning products and the development of a team mental model of the mission.

4.1.4 Defining specific behaviours

The selection of specific measures involves defining the metrics that will be used to assess human performance within each of the categories and constructs defined in the previous step. This will mean taking a major step in translating constructs into operational definitions that permit the application of an assessment methodology. For example, in mission execution, one aspect of local situation awareness during an air-to-air engagement may be the selection and execution of appropriate tactics for the situation at hand. Associated performance measures might therefore involve: accuracy in selecting and sorting targets, accuracy in adopting appropriate tactic, appropriate calculations of trajectories, time exposed to enemy weapons etc.

In most cases, it will not be possible to sample, observe or measure all critical aspects of performance, instead measurement should focus on areas that are likely to yield data that are sensitive, valid and reliable. By sensitive, we mean that they reflect the magnitude of performance variation of interest and respond to changes in variables that influence performance. By valid, we mean that they faithfully represent the category of behaviour of interest and by reliable, we mean that the measures are accurate and relatively stable.

4.1.5 Selection of measures and procedures

This step addresses specifically the issue of "how" to measure and involves the development of the measuring instruments, constructing operational definitions and determining the logistics of data measurement (e.g. real time, live-captured data playback and post event analysis). This analysis requires a detailed knowledge of the technologies and facilities available as well as the personnel resources required to implement the procedures.

For purposes of the present contract, our knowledge in some of these respects is limited because of a lack of direct exposure and access to field sites where data collection might proceed. Therefore, we have made our best guess based upon practices adopted at USAFRL and our current understanding of what is logistically feasible.

4.1.6 Defining measurement Conditions

In this process the time and place of the measurement administration are to be decided. Factors to be considered are the generalisability of the context and logistical feasibility of being able to collect the required data. Thus, a major question in the present context, but beyond the scope of this study to address, concerns the assumption that data collected in DMT contexts are applicable to actual live flying training. One particular issue of relevance to the present study is to ensure that the timing of data collection is matched to some relevant behavioural process that will be sensitive to the measures. Thus, in the application of Pathfinder techniques to assess changes in knowledge structures, there should be an appropriate selection of a training environment (or specific training course) where a significant change in such structures is supposed to occur. Our recent briefing of the Tri-Nation DMT trial in February suggests that it may be difficult to identify the operationally relevant knowledge structures that would have changed in the trial participants over the course of a few days.

4.1.7 Collect data

The individuals responsible for data collection and the time and place are established. In addition the individuals who will be the subject of the data collection are identified. This will require careful co-ordination among researchers, contractors, aircrews, analysts and AF administrators and trainers. Extensive preparation and care are required in this step to ensure that the process runs smoothly and that backup-up plans for the inevitable field contingencies are put in place.

4.1.8 Analyse data

This responsibility will normally fall to the contractor responsible for administering the study and will be guided by the scientific or operational questions of interest. A major challenge in many similar measurement contexts is to obtain enough data to conduct the usual statistical analyses and to ensure that questions of interest can be answered with some semblance of confidence. This usually means that several cohorts of participants may need to be assessed in a series of trials in order to generate sufficient data to ensure quality of the analysis, and hence represents a significant long term R&D investment by the participating parties. In evaluating performance data with respect to operational

implications, it will be generally desirable to establish criterion-referenced standards by SMEs against which the behaviour can be assessed as being acceptable, unacceptable, superior etc. Such standards would be particularly useful in evaluating qualitative performance in areas such as engagement quality, threat avoidance, terrain use and fuel/energy management.

Within the tasking of the current contract, the collection and analysis of data are not to be dealt with specifically. However, these issues are addressed in the recommendations for a Pathfinder trial plan in Annex B.

THIS PAGE INTENTIONALLY LEFT BLANK

5. Measurement Framework

5.1 Reasons for measurement in present context

Simulation has a solid history of training flying skills and is used by commercial aviation companies for a large proportion of their training. As such, flying skills are not of primary importance in the current study. Rather, given the DMT emphasis, of more interest are MOPs and MOEs pertaining to critical mission elements that would historically have been trained by flying with other pilots, equipment and nationalities.

Learning to fly effectively as part of a larger team requires an appreciation of the way others utilise their equipment and the resources at their disposal. This process implies the development of a mental model concerned with the other pilots, and possibly also the alteration of a mission-specific mental model to accommodate the actions of other aircraft. This familiarity can be accounted for in the planning process and imparted during the briefing. Likewise, any deficiencies during mission execution can be discussed during the debrief with the intention of affecting learning. Finally, progressing from training as an individual to working as an effective member of a larger team will necessitate an alteration in the components of Situation Awareness (SA) attended to by a pilot. From knowing about where the ground is, where threats are, and the state of aircraft and weapon systems, the pilot has to consider the positions of other members of friendly forces and the resources they can contribute.

In summary, the areas chosen for development of MOPs and MOEs were:

- Mission Planning (including briefing);
- Mission Debriefing;
- Situation Awareness;
- Mental Models.

The next sections investigate MOPs to assess Mission Planning, Mission Debriefing, Situation Awareness and Mental Models. Previous work by USAFRL has hierarchically decomposed mission planning into 6 phases, over 100 functions, 2000 information groups and more than 6000 information elements (Spiker and Nullmeyer, 1995). For this work, a simplified decomposition was made, based on the literature and on discussions with SMEs. This model breaks the mission events into behaviour domains, behaviour categories and specific behaviours, and specifies whether objective measures, behaviourally anchored rating scales, or Pathfinder should be used for measurement.

5.2 The DMT domain

Within the context of this project, and the DMT programme in general, the measurement areas are mission planning, mission performance, mission debriefing, situation awareness and mental model development. These measurement concepts have been described above.

One important factor to consider when deciding what to measure is what level of detail should be measured. This project aims to develop generic MOPs that should be applicable to any DMT simulation. This means that what is measured should either be applicable in all situations, or else it should be described at a coarse level of detail such that it includes specific tasks that apply to a variety

of missions, objectives or weapons platforms. Where possible, the specific items to measure are described below.

This approach is in contrast to many measurement schemes developed at USAFRL (e.g. Bennett, Schreiber and Andrews, 2002; Gentner, Cunningham and Bennett, 1998; Spiker et al, 1996). These have been developed for specific missions, specific training objectives and specific weapons systems and exploit the advanced data capture capability in the USAFRL simulators (Schreiber et al, 2002). When the Canadian Forces have such a data capture facility, it is expected that more detailed assessment of training can be undertaken. This report provides a range of objective performance data that could be collected with such a capability.

5.3 Planning, mission execution and debriefing

Spiker et al (1996) developed a conceptual model of a mission that can be applied to most weapons systems. This detailed model serves the detailed measures applied at USAFRL. This model serves as the basis for a more generic model to be used in Canadian DMT. The generic model borrows from work done by TopACES (2003) and is presented in Figure 1.

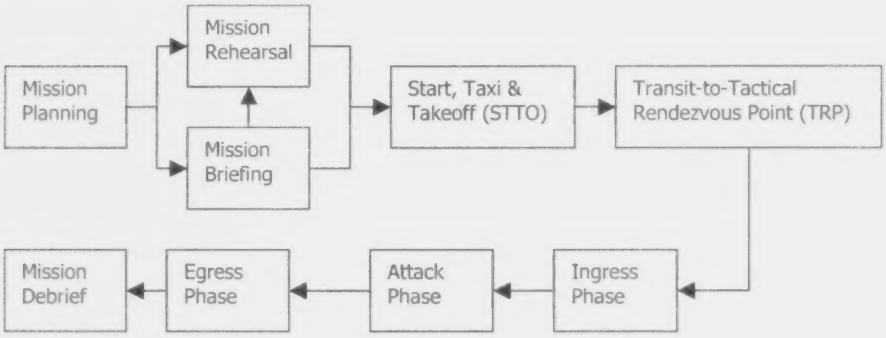

Figure 1: Conceptual Mission Model

The conceptualised mission model presented above is representative of the typical air missions simulated in a DMT environment. Due to the complex nature of air operations, significant time is spent planning, briefing, and even rehearsing the mission before it is flown. Once this planning phase of the mission is completed, time is spent on the ground: accepting, checking, starting, and taxiing the aircraft prior to takeoff. It is during this start, taxi, takeoff (STTO) phase that all aircraft systems, sensors, and weapons are verified. Once airborne, the crew follows a predetermined routing over

friendly territory to a common mission reference point, referred to as the tactical reference point (TRP).[2] Each independent formation participating in the mission is assigned a time, altitude, and speed to cross the TRP by the mission commander. It is with the use of individual crossing times over a common point that the mission participants fall into their proper order prior to entering hostile territory. Once assembled, the package of aircraft then flows towards their individual objectives during the ingress phase. The phase that follows, referred to in Figure 1 as the Attack Phase, represents that portion of the mission where the formation or crew attempts to achieve their primary mission objective(s). While the term Attack Phase refers to the typical strike fighter mission objective of conducting a successful weapons delivery on a surface target, it may also represent the establishing of a combat air patrol (CAP) orbit for the air-to-air fighter formation; the tactical delivery of troops or materiel by a transport aircraft; the insertion or extraction of troops into a landing zone by a tactical helicopter crew; or the attack run on a submarine by an anti-submarine warfare (ASW) aircraft. Following the completion of this critical mission phase, formations regroup and egress hostile territory in accordance with the mission commander's plan. Once safely behind friendly lines, the more benign RTB (Return To Base) portion of the mission is carried out ending in the safe recovery of the aircraft at home base. Having completed flight duties, crews review and study the mission during the debrief phase. This involves the accurate reconstruction of mission events, including a quantitative and subjective assessment of both crew technical and interactive (i.e. CRM) performance. An analysis of relative mission success is also carried out using the stated mission and training objectives as benchmarks. Finally, a summation of the significant lessons learned during the mission is provided by the mission lead conducting the debrief. It is emphasized that while the aircraft platform and specific mission objectives may differ from one scenario to another; the phases of the conceptual mission model described above do not, since they are inherent to air operations. The majority of research on the subject of MOPs in a DMT environment to date has focussed on the fighter mission. Accordingly, several of the examples provided in the BARS and OP descriptions in Annex A are set within the context of a fighter air-to-air mission. For example, specific references to threat reactions, aircraft manoeuvring, contact detection using onboard sensors, and weapons employment are all directly related to the tactical employment of fighter aircraft. While this is the case, it should be noted that many of the examples provided are also relevant to the generic air mission. Regardless of the platform being flown, inherent in military aviation is the need to defend against enemy threats, perform aggressive aircraft manoeuvres to accomplish tactical objectives, and employ air-to-air and air-to-surface ordnance. It is within this framework that the specific MOPs identified in this report should be considered.

5.4 The behaviour domain

5.4.1 Mission planning

Reviewing the literature, it became apparent that the elements of detailed assessment approaches to mission planning could fall into three broad categories (or domains). For instance, Spiker and Nullmeyer (1995) proposed an approach that considered component sub-plans, crew transformation and the overall mission plan. The detail to which these elements were decomposed suggested that they could be re-categorized into:

- Detailed planning;

[2] The TRP is also referred to as the time reference point. The two terms are often used interchangeably.

- Planning products.
- Briefings;

These three elements would encompass the processes by which crew transformation would be achieved, overall mission plan and component subplans would be made, and all parties would be appropriately tasked. The inclusion of planning products also allows the assessment of quality and comprehensiveness of the mission planning and briefing activities.

One further element was included for assessment at the planning stage: crew's mental model of the mission. By measuring this at the outset it is possible to quantify the changes in knowledge structure attributable to a period of DMT through comparison with a similar measurement made at the end of training.

5.4.2 Mission execution

Most mission execution measurements focus on the achievement of mission objectives. However, training to meet typical mission objectives need not occur in a DMT environment. To reflect the added value of DMT it was necessary to identify a set of behaviour domains that acknowledged the classic measures of mission execution, but broadened the category to include assessment of training effects that are more typically seen in DMT than non-networked simulators.

Following the lead of Spiker et al (1996), it was decided to adopt the behaviour domains of mission conduct and goal accomplishment. This is a framework that accommodates the traditional measurement of training missions (i.e. whether the objectives are met), but also covers the manner in which those objectives were met e.g. were resources effectively used, were communications appropriate and was a good split of attention maintained between the various pilot demands. An important additional element concerned the relationship between the mission plan and mission execution. This was subsumed within mission conduct.

Two other important behaviour domains were included: Situation Awareness and mental models. Situation Awareness has long been acknowledged as a behaviour domain critical to mission performance (e.g. Endsley, 1988). The addition of mental models as a measurement objective at this stage in training is intended to capture the 'after' phase of training in order to quantify changes in knowledge structures.

5.4.3 Mission debriefing

The final mission phase was debriefing, and there was only one behaviour domain selected for this: briefings. This reflected a perception in the CRM literature that the manner in which the crew interacts , often following the lead of the mission commander, is central to the success or failure of a debrief. Other behaviour domains, such as debriefing tool use, could have been included, but it was felt that the effectiveness of tool use depends heavily upon the intracrew dynamic.

5.5 The specific constructs

5.5.1 Mission planning

For the most part, Mission Planning performance has been assessed by SMEs either confirming pre-identified behaviours or tasks were completed (using yes/not checklists), or by rating the quality of those behaviours or tasks. In both cases, archetypal behaviours or features were defined before

beginning the assessment. Using this approach, Bergondy et al (1998) found that Mission Planning accounted for well over half of mission debriefing content for naval aviators.

Spiker, Nullmeyer and Tourville (2000) devised a set of 5 behavioural processes that could be easily measured (this measurement set was based on Special-Operations Forces):

- Function Allocation/Crew Resource Management (CRM) – division of crew responsibilities so that workload is distributed appropriately;
- Tactics Employment – analytic activities necessary to avoid or minimise threat detection or exposure, and to coordinate complex mission events;
- Situation Awareness – maintaining an accurate mental picture of mission events as they unfold over time and space;
- Command-Control-Communications – activities required to involve external parties in the mission, maintain communications with these parties, and control the sequence of mission events according to the mission execution plan;
- Time management – the ability to employ and manage limited time resources so that critical tasks are not omitted and all tasks receive sufficient time to be performed correctly.

Data collection for these processes was focused on a form that included a customised set of questions, yes/no checklist items, space for recording notable behaviours and a five-point rating scale for the crew for each process in each mission phase. Other checklists included ones to assess mission performance at each phase. During studies using this measurement procedure (Spiker, Nullmeyer and Tourville, 2001), a correlation of .78 was found between mission planning and mission performance. This finding was largely attributable to significant correlations between SA and Time Management with mission performance, and conversely, between mission planning and the low-level (high workload) portion of the mission. These findings seem to indicate that, assuming moderate levels of workload, quality of mission planning does not exert a large influence on mission performance. However, during periods of higher workload, mission planning is the foundation upon which good mission performance is built.

For present purposes, it was decided to drop situation awareness from measurement of mission planning behaviours. This was because it was considered that planning laid a foundation for the process of creating and maintaining situation awareness, but that situation awareness itself was a dynamic process that was more apparent during the actual mission. Further, the specific behaviours subsumed under SA appeared to go beyond the usual conceptualisation of SA of an integrated "mental picture" of a highly dynamic situation.

The remaining topics mirror the five behavioural processes identified by Spiker, Nullmeyer and Tourville (2000), with the addition of Planning Products, General Effectiveness and Knowledge Structures. Planning products were identified by Spiker and Nullmeyer (1995) as good indicators of mission planning performance; the more (relevant) planning products produced implies that more consideration has been given to the variety of possible ways to conduct the mission. General effectiveness serves two purposes: a quasi-validation of the other sub-scales (i.e. good general effectiveness should be associated with good scores on other sub-scales) and a unitary score of planning performance.

The behaviour category associated with mental models is simply the description of the knowledge structures possessed by aircrew before and after DMT, compared with those knowledge structures possessed by instructors (or other identified experts).

Behaviour Domains	Behaviour Categories
Detailed Planning	Tactics
	Time Management
	CRM-Function Allocation
Planning Products	
Briefings	Communication
	General effectiveness
Mental Model of mission	Knowledge Structures

Table 1: Mapping of mission planning behaviour categories to behaviour domains

5.5.2 Mission execution

Mission execution has been measured along many dimensions and using many methods. With the exception of SA, this project has not focused on traditional measures of mission performance (e.g. firing success, navigation accuracy, time, etc.). Instead, measures of mission performance have been reviewed for the extent to which they quantify the benefits of using simulation and DMT for training purposes, and for any lessons that can be applied to the measures of mission planning, briefing, debriefing and SA that are being developed for this project.

The USAF has done a great deal of work defining the Knowledge, Skills and Experiences (KSEs) required by aircrew. This provides the double benefit of outlining training requirements and assessment criteria. The USAF task list (1998) hierarchically breaks down the roles and tasks of aircrew and assigns them standard measures of performance. High level units of measurement are assigned (e.g. degree, percent, number) to complete a performance measurement framework that is applicable across the spectrum of USAF equipment platforms. The task list forms the basis of the C-130 Mission Performance Worksheet (2002). This worksheet divides a mission into a preparation, an execution, and a debrief phase. Each phase is characterised by a number of defined elements which are assessed on a five-point scale. The task list also forms the basis of USAFRL work to evaluate the effectiveness of DMT (Gentner et al, 1999b; Gentner, Cunningham and Bennett, 1998).

A close Canadian equivalent to this is the table of contents from the CF-18 'learn to fly' manual (2000). These headings form a first level breakdown of performance measurement elements. Further hierarchical decomposition and assignment of units of measurement would be required in order to develop an instrument that could be used by flight instructors/assessors. This may have already been done in the Canadian Air Force's 'Training Task Analysis', which was not available to be reviewed during this project. During DMT trials, the trial timetable and ATOs (e.g. Trial VirtEgo, 2001) could also be used to develop a detailed performance measurement framework.

Similarly based on a detailed decomposition of the aircrew task, the USAFRL have also developed Mission Essential Competencies (MECs: Bennett, Schreiber and Andrews, 2002). As the name suggests, these MECs are specific to particular mission types. MECs are higher-order individual, team, and inter-team competencies that a fully prepared pilot, crew or flight requires for successful mission completion. Each MEC is then further defined by identifying a purpose, and a beginning and ending point. The MEC framework also includes some (more general) supporting competencies, and lists of knowledge and skills. In addition, a list of experiences through which these knowledge, skills and competencies can be learned has been generated. MECs are not abstract knowledge or general

skills. They are demonstrated to a degree of competence in the context of an actual mission or high-fidelity simulated mission. Using competencies to structure training and assessment is an approach that has also been adopted by Naval Air Wings (Bergondy et al, 1998).

Consideration of this literature led to the adoption of the following behaviour categories to match the behaviour domains. Note that CRM, tactical awareness, mission/goal awareness and flying skills are not associated with specific behaviour domains as a reflection of their overriding influence on mission execution. Each behaviour category is broken down into greater detail in the next section, which lists the specific behaviours associated with each category.

Behaviour Domains	Behaviour Categories
Mission conduct	Plan compliance
	Communications
	Aircraft handling/control
Situation awareness	System awareness
	CRM
	Tactical awareness
	Mission/goal awareness
	Flying skills
Goal Accomplishment	Achievement of objectives
	General effectiveness
Mental Model	Knowledge structures

Table 2: Mapping of mission execution behaviour categories to behaviour domains

5.5.3 Mission debriefing

Only one reference (Morrison and Meliza, 1999) explicitly dealt with the debriefing process, and this provided guidance in conducting an after-action review rather than how to assess and train debriefing skill. However, it is clear that debriefing shares many of the features of mission briefing. This, and the fact that debriefing is often considered the end result of a training opportunity and thus implies assessment, has meant that debriefing has not been the subject of specific investigations to develop MOPs.

Two papers deal with debriefing obliquely, but usually within the context of CRM (Spiker et al, 1999; Nullmeyer and Spiker, 2002). Air Force Instruction (AFI) 11-290 (1 July 1998) defines 6 CRM categories, including Mission Evaluation which includes post-mission debriefing. This AFI also requires each Wing to collect proficiency-related data in accordance with the structure afforded by the 6 categories.

One proposed method of gathering this information (Spiker et al, 1999) broke each CRM dimension down into 3 observable elements, specific to each mission phase (mission preparation, leg 1, leg 2, mission debrief). A SME noted the presence or absence of these behaviours and rated individuals and the whole crew on their general performance of that CRM category using a 5-point scale. Another

observer used a behaviourally-anchored 5-point scale to rate crew performance during each mission phase.

The Spiker et al (1999) study found that mission debriefing did not significantly correlate with mission performance, but did make recommendations for the debriefing process. These recommendations were that training environments should be reviewed to ensure: 1) crews have leaders who take individual responsibility; 2) a non-threatening (non-blame) atmosphere is maintained; and 3) all instructors/evaluators become 'facilitatory qualified'.

Nullmeyer and Spiker (2002) included mission debriefing amongst the mission preparation behaviours they attempted to find in the archival data. They did not find any detail in training records that explicitly dealt with debriefing performance and behaviour. Further, in relating the archival data to flight safety reports, they did not find mission debriefing cited as a causal or contributing factor to any incident.

As noted in above, it was felt that briefings encompassed the most valuable debriefing behaviours. Communications and crew CRM obviously reflect the interpersonal skill aspect of debriefing, but the degree to which mission accomplishment and the crew's technical performance are critically appraised also depend upon the manner in which the crew relate to each other and leadership provided by the commander. Another important objective of mission debriefing is to learn through the experience and feed this learning forward into future missions. This led to the inclusion of lessons learned, to reflect the need for debriefs to assist in developing performance. General effectiveness is included to capture overall performance and to validate scores on other sub-scales.

Behaviour Domains	Behaviour Categories
Briefings	Communication
	Mission accomplishment
	Crew CRM
	Crew Technical
	Lessons Learned
	General effectiveness

Table 3: Mapping of mission debriefing behaviour categories to behaviour domains

5.6 The specific behaviours to measure

5.6.1 Mission planning

Nullmeyer and Spiker have been the most active in investigating the benefits of Mission Planning on Mission Performance (in the specific context of simulation exercises). Spiker, Nullmeyer and Tourville (2000) referred to a set of 10 core principles that underpin the planning process and seem to characterise the best plans (after Spiker and Campbell, 1994):

Plan from the target area backward;	Become familiar with the terrain around the target area and the visibility conditions that will exist at the time-on-target to determine constraints on possible tactics.
Plan from general to specific;	Consider possible aircraft platforms and general tactics before specific avionics configurations and specific tactics.
Plan from high to low;	Determine timing, communications and fuel requirements for the high altitude part of the mission first, since low-level route construction will be constrained by these factors.
Plan from big to small;	Use a small-scale map to visualise the enroute return portion of the mission. Use a large-scale map to visualise the area of operations and target area.
Plan from qualitative to quantitative;	Use qualitative rules of thumb during the general planning phase, when mission feasibility and tactics are being established, before working with precise numbers.
Build slack into the calculations;	Allow a 10% 'fudge factor' in aircraft fuel flow calculations. Assume the onboard load will be 10% higher to accommodate unexpected passengers and/or cargo.
Always plan for the worst case;	There should be no aspect of the plan that depends on good luck or optimal conditions for success.
Fly as high as the threat dictates but as low as the terrain allows;	Make sure the planned vertical clearances are not overly low, since crew workload will go up. Make sure that set clearances are not overly high, since large changes in altitude above ground level will be required to avoid threats, which also increases crew workload.
Keep the plan as simple as possible;	If the basics of the plan can't be briefed in ten minutes, then it is too complicated.
Time is the true enemy.	Do not use up all the allowable time during general planning. Be sure to save a substantial portion of the time to complete time-consuming, detailed planning (e.g. precise fuel flow calculations, specific waypoint determinations, etc.).

Table 4: Core Principles of Planning

Spiker and Nullmeyer also considered Special Operations Forces (1995) and identified 42 measures of mission planning effectiveness. In Spiker, Nullmeyer and Tourville (2000), 12 of these behaviours were chosen (due to their relevance to the particulars of their study) and crews were scored on the extent to which those behaviours were observed during planning activities. Observers could rate each behaviour as either -1 (a negative example of the behaviour) or +1 (a positive example of the behaviour. Therefore crew scores could range from -12 to +12. This simple measure of planning effectiveness resulted in a significant correlation with mission performance (.71). It is not possible to list all 42 behaviours identified originally, due to the security classification of the work, although 12 behaviours with exemplars are listed in Table 5 for illustrative purposes.

Mission Preparation Behaviours	Exemplars
All planning personnel are effectively utilised	Aircraft Commander (AC) asked all crewmembers for 'what you need to do your job' and then got it for them
Establish timeline for managing the planning process	AC told crewmembers when they had to be completed with their planning tasks in time for the crew briefing
Precise times are determined for accomplishing the key mission events	Planned times and routes backward from a specific point. Determined optimal take-off time on this basis
High quality crew briefings are given during various stages of planning	After each crewmember briefs, the AC adds final comments for the crew's consideration
Planning crew achieves an in-depth awareness of threat capabilities along the route	To avoid threats crew planned to fly very low altitude, terrain mask, and high speed (as necessary) manoeuvring
The plan is developed to an appropriate level of detail	Crew members directed to devise evasion plan of action
All information sources are checked for recency	AC asked when intel had last been updated
Information is cross-checked for accuracy and the plan's assumptions are aggressively questioned	AC questions assumptions in each crewmember's plan
Ground team and support asset requirements are incorporated into overall plan	AC modifies plan to incorporate considerations of (e.g.) helicopters for transload
Mission essential equipment is well thought out and incorporated into plan	Crew listed the minimum equipment needed to accomplish the mission, such as INS, chaff, flares, etc.
Planning assumptions are subject to extensive 'what iffing'	Crew planned to 'bump up' their airspeed if they encountered threats during certain mission phase
Planners incorporate their real world experiences into the planning process	Crewmembers related their own experiences in the area of operations as they developed the execution plan

Table 5: Examples of Effective Mission Preparation Behaviours (from Spiker, Nullmeyer and Tourville, 2000)

It is apparent from the behaviours listed above that a strong CRM skill set will enhance the probability of effective mission preparation behaviours, which, in turn, will increase the probability of effective mission planning products.

The Canadian Air Force has also defined a set of 8 principles of planning ('Learn to fly', 2000 chapter 11, 11-5 to 11-9). These 8 principles form the mnemonic 'tight wad lady' (TTWODLDE) and stand for:

- Target;
- Terrain;
- Weather;
- Ordnance;
- Defences;
- Lateral Support;
- Delivery;
- Egress.

These principles of planning seem to focus on maximising SA, which in turn supports tactics employment. In combination with the principles and behaviours defined by Nullmeyer, 'TTWODLDE' may present a useful framework within which to assess planning behaviour.

Further interesting findings arose from the Coalition Mission Training Research (CMTR) exercise in February, 2003. Pilots at all facilities (US, UK and Canada) stated that the most value from DMT would be gained through focusing on mission planning and briefing behaviours. In a summary of the exercise, Grant (2003) stated that, for the most part, individual sites engaged in local planning (that is to say, they did not collaborate with other sites). When interaction between sites was pursued, it was usually the case that the mission commander provided a briefing, and sought confirmation of his plans from the other sites. Thus planning was not collaborative in the CMTR DMT exercise. However, this finding does suggest a package of measures focusing on the nature of planning processes (as opposed to the content). For instance, the number of open versus closed questions or the number of information gathering/collaborative questions versus confirmation questions could be counted. This also reduces the need for SMEs or the use of questionnaires as a novice analyst could successful code the data from videotape with a small amount of training.

Ultimately, the following list of specific behaviours of interest during mission planning activities was distilled from the literature:

Behaviour Domains	Behaviour Categories	Specific Behaviours
Detailed Planning	Tactics	Task understanding ROE understanding Route review/analysis Tactical effectiveness of plan Factors considered in plan COA considered Decision quality/timeliness Use of resources Creation of extra materials
	Time Management	Time appreciation Time required Efficiency
	CRM-Function Allocation	Clarity and assignment of team roles
Planning Products		Quality of: Fuel plan TOLD Communications Mission data
Briefings	Communication	Detail Participation Comprehensiveness Time required
	General effectiveness	Degree of instructor intervention
Mental Model of mission	Knowledge Structures	Change in KS resulting from training

Table 6: Mapping of mission planning specific behaviours to behaviour categories

5.6.2 Mission execution

Castillo et al (2002) used the MEC approach by developing a Situational Judgement Inventory (SJI) and a Job Knowledge Inventory (JKI). These two inventories are databases respectively describing the likely situations aircrew are likely to face and the knowledge that needs to be obtained before the situations can be successfully overcome. SJI and JKI are the foundation of a vignette-based method of knowledge assessment termed Situation Assessment and Action Selection (SAAS). Archetypal solutions are developed for each vignette and student's answers are scored relative to that. With greater experience of measurement of DMT, this may become an approach worth pursuing.

The performance measurement approaches described above focus on behaviours exhibited by the aircrew. This measurement approach tends to be labour intensive, but is necessary to make assessments of qualitative aspects of task performance or tasks that are not amenable to automated data collection. Another alternative is the use of event or scenario-based measures of performance (Gentner et al, 1999; Dwyer et al, 1999). The use of event-based measures of performance can permit the use of automated performance tracking systems (Schreiber, Watz and Bennett, 2002) and leads to greater inter-rater reliability if SMEs are used for assessment (Dwyer et al, 1999).

Event-based training is an instructional approach that systematically structures training in an efficient manner by tightly linking learning objectives, exercise design, performance measurement and feedback. Thus, event-based training builds on the USAF task list and MEC approach described above. Event-based training requires that trigger events be identified for each learning objective. These events are the stimulus conditions and cues that appear in the exercise and require a response by the participants. This allows the participants to demonstrate their ability to perform the tasks associated with the learning objectives, and allows SMEs (or the system) to assess the response against predefined criteria and provide targeted feedback.

The consideration of behaviourally and event-based measures of performance led to the compilation of the following list of specific behaviours of interest with respect to mission execution. These behaviours represent generic features of performance that can be described and assessed by SMEs observing a DMT mission.

Behaviour Domains	Behaviour Categories	Specific Behaviours
Mission conduct	Plan compliance	Navigation accuracy Time control Communications required Compensation/adjustment Loss of separation incidents
	Communications	Speed/accuracy Terminology Necessity Information content
	Aircraft handling/control	Piloting skills (e.g accuracy in heading, speed, altitude, clearances)
Situation awareness	System awareness	Checklist performance
		Sensors, status indicators
	CRM	Crew awareness and communication
	Tactical awareness	Contact detection(speed,accy) Spatial awareness Co-ordination COA
	Mission/goal awareness	Resumption of plan after engagement Detects changes in mission picture Changes/updates plan re changes in mission picture Communicates with crew
	Flying skills	Engagement skills Weaponry skills Threat exposure Role discipline
Goal Accomplishment	Achievement of objectives	Completeness, survival
	General effectiveness	Degree of instructor intervention
Mental Model	Knowledge structures	Team mental model/co-ordination/CRM

Table 7: Mapping of mission execution specific behaviours to behaviour categories

5.6.3 Mission debriefing

Intuitively, it seems sensible to conduct a comprehensive debrief. Using the model of mission events described in section 5.3 and the Canadian Air Force mnemonic TTWODLDE, one can develop a checklist of debrief topics. In this checklist, each of TTWODLDE would be considered for each mission stage (STTO, TRP, Ingress, Attack, and Egress). The SME observing the debrief could then assess the debrief for its comprehensiveness.

The content of debriefs is less structured than can be expected for planning and mission execution. This could be because the debrief focuses on learning lessons from a mission, and if the mission was successful, the perception is that there is less to learn. As a result, generic measures of debrief performance are harder to develop. The measures listed below largely focus on the amount of time spent discussing facets of mission performance. Since the exact content of the debrief will depend upon the manner in which the mission was conducted, the other sub-scales rely on SME perceptions of completeness.

Based on a combination of the measures outlined in this section, and other measurement methods described in other sections, a number of specific behaviours are proposed below.

Behaviour Domains	Behaviour Categories	Specific Behaviours
Briefings	Communication	Detail, Participation, Comprehensiveness, Time spent
	Mission accomplishment	Time spent Completeness
	Crew CRM	Time spent Completeness
	Crew Technical	Time spent Completeness
	Lessons Learned	Time spent Relevancy
	General effectiveness	Degree of instructor intervention

Table 8: Mapping of mission debriefing specific behaviours to behaviour categories

6. Measures of Performance

6.1 Summary literature review

As noted previously, a great deal of literature was reviewed for this contract. Although some of the literature discussed automatic collection of objective data, most discussed researcher- or SME-mediated collection of data and assessment of performance. This reflects the contract's emphasis on mission planning and debriefing, as well as the requirement to develop generic MOPs for application during DMT scenarios.

The large proportion of measures reported in the literature took the form of Behaviourally-Anchored Rating Scales (BARS). These BARS attempt to ensure validity and reliability by providing descriptions of the types of behaviours that represent good, expected, and even poor performance on a particular measurement dimension of interest. BARS represent expansions of the Mission Essential Competencies (MECs) developed by USAFRL, which themselves are expansions of the high level Air Force Task List (AFTL). The AFTL and the MECs required further decomposition in order to become useful in the assessment of aircrew performance because they represent Measures Of Effectiveness (MOEs) as opposed to Measures Of Performance (MOPs).

Measures of Effectiveness have focused on the efficiency or capability of a system to achieve some goal state, or on the ability of a system to mediate in the achievement of a goal state. This goal state is often an abstraction of some combination of tasks and is unlikely to be 'active' in that it is more likely to be a state (of mind or readiness for example) as opposed to the successful completion of a task or process according to some criteria. For example, in the context of the current project, the MOE is intended to determine how well DMT mediates in the development of knowledge, skills and aptitudes of aircrew. This can be broken down to questions of the following sort:

- Does DMT result in adequate improvement of mental models?
- Does DMT result in adequate improvement of SA ability?
- Does DMT result in adequate improvement to planning, briefing and debriefing skills?

These questions are asked at a 'macro' level of detail. Well-constructed MOEs will also be subject to pre-set success criteria, often representing the level of effectiveness demonstrated by the system to be replaced or changed.

Measures of Performance have tended to focus on the execution of a task or process, and are stated as characteristics of what is accomplished (e.g. speed, accuracy). MOPs are usually direct output measures of task or process execution and, in combination, are often predictors of the effectiveness of a system. In the context of the current project, MOPs are intended to determine how well aircrew achieve discrete events or how well they exhibit certain behaviours. MOPs might include (but are not limited to):

- What was the reaction time to <<some event>>?
- What was the time on task?
- Did they notice <<some SA probe>>?
- Did they exhibit <<some behaviour>>?
- Did they make contingency plans for <<some eventuality>>?

These questions tend to be asked at a more 'micro' level of detail. Well-constructed MOPs are usually written with some stated threshold, above or below which will be considered success or failure.

In most of the studies reviewed, BARS are used to provide a numerical rating of performance, leading to subjectively-derived quantitative data. This is distinct from objectively-derived quantitative data, which represents a measurement of some observable feature of the data that does not itself require any interpretation and analysis. Other subjectively-derived data include analyses of training records. Although the reviews result in counts of good outcomes, bad outcomes, etc., the data is subjectively-derived because it initially relies on instructor assessments of performance.

Pathfinder analysis also results in subjectively-derived quantitative data. The subjective element arises from two sources: the selection of candidate concepts for rating and the rating for each relationship provided by the aircrew themselves. Since objectively-derived measurement of knowledge structures seem unlikely, this subjectively-derived approach will continue to be the most accurate method of describing knowledge structures for some time yet.

The scope of this project allowed not only for the development of subjectively-derived MOPs, but also MOPs that result from a more objective process. A number of objective performance metrics have been developed for this contract. The development of these objective MOPs is based on discussions with SMEs, familiarisation visits to USAFRL and DRDC, and evolutions of the MOPs described in the literature.

6.2 Recommendations for measures

Based upon the literature reviewed and analysed in previous sections, we now outline our recommendations for the specific methodologies to be used for the collection of MOP data. As has been outlined above, there are three basic strategies for measuring performance, each of which depends upon the activity to be studied, the aspect of performance of interest and the logistical constraints in conducting the mission. The three approaches are:

- Objective performance data
- Behaviourally anchored rating scales (BARS) scored by SMEs
- Pathfinder

The first two of these measurement strategies can be applied to most of the functions and perceptual/cognitive processes required of crews in mission planning, execution and debriefing. Pathfinder analysis is exclusively to be used for the assessment of the crew knowledge structures.

6.3 Objective performance data

In general, this approach is based upon directly measuring the processes and products of the category behaviour of interest. The processes may be actual operational procedures, functions to be performed or the underlying human processes. If we take the planning process as an example, briefing a team and the accuracy with which it is conducted would be an example of an operational procedure and associated measure; time management during planning would be an example of how effectively the planning process is organised. In addition, the products (navigation plan, communication plan etc.) that are produced as part of the planning process can be assessed for completeness and accuracy (when ground truth is available).

Objective data are represented in terms of one of four measures: response time, duration, accuracy (error) and frequency. It is assumed that readers of this report are familiar with such measures and that there is no need to provide explanation or examples of each. One major limitation in the collection of accuracy and error data may be the limited number of events over which data can be aggregated to

generate a representative statistic. Military opportunities to conduct measurement trials may be limited, the number of participants few and there may be only one or two events over which to sample the behaviour of interest. Hence, reliable and accurate estimates of error rate will be difficult to obtain.

Ideally objective performance may be assessed against a quantified performance standard. In practice, this is rarely available. Therefore, performance data may tend to be used on a comparative basis, or they could be used to inform subjective opinions about what values correspond to poor, standard or exemplary performance.

While objective performance data may seem to be the "gold standard" of measurement approaches, they are not immune to challenges to validity and reliability. The validity of an objective measure may frequently relate to its relationship to overall task performance. Thus, although a process may be accurately measured and quantified this does not address the issue of the importance of the process to the behaviour goals in question. A process that might be easily and reliably measured, such as response time to select a menu item in an avionics display, may only comprise a small proportion of the overall variance for the larger task at hand. In the absence of a detailed, functional model, which identifies critical tasks of the process of interest, the measurement strategy will have to rely on intuition and reasoning based upon SME input and an analysis of any relevant literature.

In the literature on the evaluation of aircrew performance that we have reviewed to date, there has not been extensive use of objective performance data. There may be several reasons for this: a lack of suitable technology to record and analyse the data of interest or insufficient resources available to perform the detailed analysis. Indeed, these factors could be seen as major barriers to using this approach, especially to analyse complex missions with multiple players that have the potential to produce high volumes of real-time data. Our own experience in analysing operational data from multi-player complex teams (Matthews, Bruyn, Keeble and Webb, 2003), suggests that for every hour of recorded data, as much as twenty person hours are required for analysis, even with a highly functional playback/analysis suite.

Thus, our recommendations for specific objective performance data in subsequent sections are based upon an identification of mission critical tasks and the practicality with which they can be measured and analysed in a DMT context.

6.4 Behaviourally anchored rating scales (BARS)

BARS have been widely used to assess human performance for a wide variety of applications and appear to be the method of choice of the USAFRL. Rating scales in general have been widely used for assessing performance (Landy and Farr, 1980) and they have broad appeal for their intuitiveness, ease of administration and versatility. Meister (1985) notes that rating scales may be used in the following ways:

- To evaluate how well a task is being performed (e.g. flying a plan precisely)
- To evaluate some quality of performance (e.g. leadership of the OC during mission execution)
- To quantify the adequacy of some feature of a system (e.g. the detail of navigational plan)
- To evaluate the effect of some condition (e.g. DMT versus live flying)
- To evaluate the output of performance (e.g. the choice of a tactical strategy)

While there are a number of approaches to the construction of ratings scales, BARS have emerged as a preferred technique because the behavioural descriptors are thought to provide the necessary anchors to enhance the precision of the rating, to standardise across observers and to screen out idiosyncrasies.

An example of such a scale to assess mission briefing is shown below.

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1. COMMUNICATION: Mission briefings: detail, participation, comprehensiveness				
Little or no mission briefing. Little or no crew participation Time wasted on irrelevant communication	Minimal mission briefing and crew participation. Most elements covered. Provides little detail on objectives	Brief contains some inconsistent mission details Few crew participate	Detailed briefing Several crew participate Many elements covered with acceptable level of detail	Highly detailed mission brief. All crew participate All mission elements covered with great detail
Look for: Does AC do all or are parts delegated? Crew members planning responsibilities defined Was key info omitted? Was it rushed? Did it ramble (unfocussed)? Does crew appear to leave briefing with understanding of mission? - with confidence? (For DMT) were there appropriate communications with remote mission participants? Note duration of briefing:_____ (mins) **Observations:**				

Figure 2: Example BARS

Clearly such an instrument has clear advantages over simply asking raters to "on a scale of 1-5 rate the quality of mission briefings". First, the scale defines the behaviour category of interest, in this case "Communication". Second, it summarises the range of communication behaviours of interest, i.e. the content detail, who participates and, the scope of coverage. Third, it provides descriptors of the performance standard that merit the different rating values. Fourth, it directs the rater to specific key behaviours.

There is an extensive literature on the validity and reliability of BARS and other subjective rating scales (see Meister, 1985 for an overview). This literature shows a large variance and discrepancy concerning the sensitivity, effectiveness, error bias of BARS and other rating schemes, largely dependent upon the behaviour being assessed, who the subjects are and who are the raters.

A number of common errors have been identified with the use of ratings scales in general.

Errors of *leniency* occur when the behaviour tends to be rated towards the upper end of the scale. USAFRL has reported such errors when instructors responsible for training are then asked to rate mission performance behaviour.

Sequential errors occur when later ratings are affected by earlier ones. The usual randomisation process that is employed in laboratories as a strategy to counter such errors is probably not feasible in a DMT environment and hence, researchers must be vigilant in seeking out where such errors may

occur. Rigorous training, followed by monitoring of raters that incorporates performance (rating) feedback may be one approach to overcoming this in the field.

Distribution errors are when ratings tend to be congregated around middle values and result in a lack of discrimination or sensitivity to changes in the observed behaviour. Again, such errors have been reported by USAFRL researchers who then used strategies of more explicit instructions and revised training procedures of raters to try to minimise such effects.

The final type of error is *inter-correlational*, whereby ratings on one behaviour may influence other behaviours being observed, a so-called halo effect. Errors of this type have also been observed in USAFRL studies, for example a crew that is observed to be highly effective in accomplishing standard tasks at the beginning of a mission, may tend to be rated high in subsequent engagement or weapons delivery phases, even when the performance does not warrant such a high rating.

Fortunately, psychologists at USAFRL, aware of such limitations and cautions, have done much of the necessary quality assurance in developing BARS for the evaluation of mission performance, during planning, execution and debriefing phases. This has included the required steps of determining critical tasks, defining the performance dimensions of the tasks, using SMEs to scale the performance of the tasks, deriving the behavioural anchors, and, in some cases, testing the psychometric properties of the instrument. Wherever possible we have borrowed from this extensive battery and adapted and evolved other similar instruments to tasks and performance dimensions that are the focus of the current work. Our confidence in the recommendation for the use of BARS is based upon clear indications of its reliability, validity and sensitivity that have been demonstrated in a number of USAFRL studies that have successfully used the tool to assess and measure the relationship between mission planning and mission performance (e.g. Thompson, Tourville, Spiker and Nullmeyer, 1996; Spiker, Nullmeyer and Tourville, 2000).

6.4.1 Issues for the administration of BARS

The following recommendations are made based upon the USAFRL experience, our own experience in administering such scales in a military context and the general literature that we have reviewed.

1. Where a new BARS scale needs to be developed, a pool of SMEs should identify the critical behaviours, define the performance indicators and scale descriptors. Mission essential competencies, training goals and standards and training task analyses may all be sources of useful information to assist this process. The scale should be pilot tested to ensure that it has the required sensitivity.
2. To avoid bias in scoring, different sets of SMEs should assess crew performance during briefing, mission execution and debriefing. Wherever possible, trainers who may have already developed opinions concerning the performance quality of trainees during prior instructional phases, should not be used to assess performance in the mission scenario of interest.
3. Ideally, SMEs doing the ratings should be given an opportunity to develop their assessment skills in a pilot test that provides standard behaviour samples and standards against which their ratings can be assessed.
4. Researchers need to be vigilant for error bias in the ratings. They may need to encourage observers to use the lower ends of the scale range, when appropriate and to be consistent in judging the behaviours according to the scale descriptors. It has been reported that trainers may be reluctant to give crews low ratings, even when the performance merited it.

6.5 Pathfinder

6.5.1 Overview and historical development

The antecedents of the development of Pathfinder can be found in attempts by cognitive psychologists and others working in artificial intelligence to find ways of formally and mathematically representing human cognitive structures. The Pathfinder methodology finds its roots in both multidimensional scaling (MDS) and cluster analysis techniques. Together with graph theory, these techniques have made substantial contributions to connectionist models of cognition and provide a novel technique for the representation of relationships that may exist with a knowledge structure.

The major step forward provided by the Pathfinder approach has been the use of empirically derived data from relatively small subject populations (compared with standard MDS and other clustering approaches) to derive graphic representations of knowledge structures.

One of the earliest applications of the model was to examine the relationship among 25 naturally occurring concepts (e.g. items such as blood, animal, plant, feathers - see Schvaneveldt and Durso, 1981). The resulting network (Pathfinder Net - abbreviated PFNET) derived from the analysis provided a means of showing graphically the inter-relatedness among the concepts.

More recently, and of relevance to the present work, the approach has been used to look at expertise, or differing levels of expertise, in the context of pilots (Schvaneveldt, Durso, Goldsmith, Breen, Cooke, Tucker and DeMaio, 1985). The ensuing PFNETS for inexperienced or experienced pilots allowed the accurate classification of the representation of a single pilot (whose experience was unknown to the researchers) to be accurately determined.

6.5.2 Details of core concepts

The core, underlying concepts that shape the Pathfinder technique are Graph Theory and network structures. Graph theory provides a method for representing an abstract network structure that is interconnected through pairs of nodes and their links, whereby the nodes represent the core concepts and the links represent the relationship. The network is represented by a graph that shows the weight associated with each link. Across individuals, similar networks imply similar knowledge structures. Combining groups with similar knowledge structures (e.g. experts) into a common network, allows detailed comparisons of the knowledge structures with other individuals or groups (e.g. novices).

Nodes are a finite set and the links are a subset of all node pairs. The specific nodes that are connected by a link are designated the endpoints of the link. The link is said to be incident to a node if the node is an endpoint of a link. The degree of a node is defined as the number of links incident to the node. The graphs of the nodes and links can be of two types - directed (digraph) or undirected. The directed graph implies directionality to the link - that is going from the first node to the second. An undirected graph has undirected links (edges) in which the nodes connected by the link are regarded as unordered.

A connected graph is said to contain a path between any two nodes. One variation of this is a tree in which there are no cyclical paths, whereas a complete graph shows all possible links.

The network links have positive real numbers that may represent weight, distance or costs; its associated graph is generally represented by deleting the weights. Thus, the graph represents the network structure, whose relational structure is provided quantitatively by the link weights.

A PFNET is obtained by first generating proximity data, which requires individuals to rate the psychological distance of things that belong together; for example, most people would rate the terms doctor and hospital as being closer together than doctor and rainbow. The proximity of two concepts

is therefore operationally defined as their rated similarity. Practically, small proximity values mean close relationships, larger values more distance or lack of relatedness.

A complete network will show each node with the weight on each link equal to the proximity of the nodes that it connects. However, because of the amount of detailed information plotted, the representation is not informative. The PFNET that is derived from the complete network only considers the most proximal links between the nodes, that is the links that represent minimum weight paths. Thus, the PFNET has an equivalent distance metric as a complete network but only represents the minimum number of links that are required to yield that distance metric.

A PFNET is characterised by two parameters r and q, which represent generalisations of standard definitions of path lengths or distances in a network. The r parameter determines how the weight of any path is computed from all the links in the path and the q parameter sets a limit on the number of permissible links in a path. For networks derived from empirical data, where the scale properties of the numbers generated cannot be assumed to be higher than an ordinal level, r is usually set to infinity. The q parameter may be varied according to a number of contextual circumstances and essentially sets the density of the links in a PFNET. For example, it may not make sense from a theoretical or cognitive perspective to allow all possible links to be computed, since this would be uninterpretable and have little explanatory value. In most of the studies reviewed, q has been set to n-1, where n represents the number of concept items of interest.

Readers are referred to Dearholt and Schvaneveldt (1990) for more detailed descriptions of the Pathfinder model and the underlying statistical properties and mathematical proofs.

6.5.3 Steps in the Pathfinder approach

Essentially all Pathfinder paradigms can be divided into three distinct phases: concept elicitation, psychological scaling and interpretation. Within these areas we have identified the typical steps and processes that are involved.

Concept Elicitation

1. Defining the measurement domain of interest.

While this may seem obvious, it cannot be approached too casually if one wants to avoid problems with the external validity of the ensuing network generated. For example, in a training context it would be desirable to limit the domain to the specific knowledge structures of the trainee that might be expected to be changed by the learning process. Since, as we shall see in the next step there are practical limits to the number of concepts selected, the initial challenge is approach the selection of the domain in a disciplined and conservative manner.

2. Selection of the concepts that are believed to be represented in the knowledge structure.

Approaches to this have varied from the informal - a group of "experts" discusses and reaches consensus on what constitutes the set, or formal - where a structured approach is taken. The structured approach may involve the selection of concepts by experts from training objectives, performance criteria, essential competencies and other formally derived representations of the domain of interest. This selection may then by refined by a number of formal processes such as ranking, sorting and other comparison techniques. Indeed, a preliminary PFNET could be derived from the initial set of concepts and these could be refined according to the item analysis that can be readily derived using a utility in the Pathfinder software. Some iterations of the process may be necessary to achieve the desired set.

3. Reducing the number of concepts to the largest number that can be practically handled by the expected subject population.

Since each concept will be compared with each other, the number of pairwise comparisons increases dramatically as the number of concepts increases, using the formula n((n-1)/2). If the subject population is one in which time pressure is not a consideration, then a large number of concepts can be selected. However, in the case of military personnel, where time to participate in "studies" must be squeezed into already busy schedules, a practical limit may be 20-25 concepts, which may take the subjects 30-45 minutes to do the pairwise comparisons.

Psychological scaling

4. Selection of study participants.

In the present context, Pathfinder is planned to be used as a tool for evaluating the change in knowledge structures that result from military training. Therefore, it would make sense to select a sample of military personnel who are at the beginning of a defined training module and to conduct the Pathfinder analysis just prior and immediately at the end of training. As a further referent for evaluating changes in knowledge structures, a sample of instructors or operational personnel, with current domain knowledge could also be selected.

5. Determination of sample size

As with most statistical techniques, improved reliability and validity are obtained with larger subject samples. However, Pathfinder has been shown to produce valid PFNET solutions with sample sizes as small as six. To date, we have come across no specific recommendations or validation procedures that have systematically related the quality, generalisability and robustness of a PFNET solution to the sample size.

In the case of studies in which the effects of training are to be evaluated on a group wide basis it makes sense to have as much homogeneity in the sample population as possible with respect to variables of influence. However, if the interest is individual differences then samples should be selected to ensure some variation along the critical variables.

For the purposes of any future study involving military personnel in a training context we would recommend trying to obtain a complete cohort of trainees for any single course offering. In the case of obtaining comparison networks from experienced personnel or trainers, where access and availability may limit the "n", we would recommend a sample size of no less than six, but preferably larger than 10 if feasible.

6. Conduct pairwise comparison of concepts

The selected sample of participants performs a pairwise ranking of all of the concepts using a five point scale. The concepts should be randomised in advance and the serial order randomised across participants. If computer technology is readily available, this task can be conducted on-line and the data captured and stored in real time, to save time and potential errors in subsequent transcription.

Experience has shown that it is important to provide some motivation to the participants to ensure that appropriate judgment is brought to bear and the task is not performed in a casual or cavalier manner.

The process of conducting the comparisons should be performed prior to and immediately following the period of training (or activity of interest). Again, some inducements may be necessary at the second procedure as participants may have little motivation left for such a task after a particularly gruelling period of training and when there is a strong desire to get away. Experience has shown that

it is unwise to have this task performed on a Friday afternoon after the final wash-up, which is the time usually made available by the military. To overcome this, it may be desirable to do the task earlier on the final day or to provide some real incentives at the end of the day to better ensure the task is completed with motivation and integrity. Instructor buy-in to the study objectives must also be sought, again to ensure that the associated tasks are treated seriously.

7. Trim data for outliers

Since it is possible that some participants will not approach the sorting process with the appropriate level of diligence or commitment, they may generate data that is inconsistent with the overall group pattern. With small data sets, this will decrease the reliability and sensitivity of the study. Therefore, it is recommended that *a priori* rules be established to trim the data for outliers. Such rules might include assessment of the lack of variation in the rankings across pairs, or rankings that are two or more standard deviations higher/lower than the group mean, or variances across ratings that are inconsistent with other group members.

Interpretation

8. Analyse and interpretation of data

The data are analysed using the Pathfinder software which runs in a PC-Dos environment. The investigator must make a decision on what value to set *q*, if the default value of n-1 is not to be used. We have not been able to find any clear guidelines or practices that will guide this selection. It has been suggested that q be varied on an ad-hoc, individual subject basis in order to generate PFNETs with similar numbers of links across subjects. Further, the graphic representation of the PFNET solution must be reviewed for layout consistency and interpretability and the user should be prepared to manually edit the location of nodes in order to improve the default output. There are no guidelines for this process and the intuition of the user will be an important factor in this process.

The PFNET software provides a broad selection of statistical properties that can be computed for the network solution including important properties such as item reliability and outliers and internal consistency and coherence. Coherence measures have been used to reflect expertise in the measurement domain, with more expert raters producing higher coherence ratings (Schvaneveldt, Tucker, Castillo and Bennet*). To compare networks across different groups (or individuals) the similarity of an individual's *ratings* at the beginning of training is compared with average expert *rating* - this is then repeated at the end of training. In addition, the change in similarity to expert ratings over the course of training can be computed by taking the difference between these similarity assessments. A similar set of procedures can also be conducted to assess the similarity of the *network* between an individual and a reference group. A more detailed review of comparison procedures is outlined in a subsequent section.

6.5.4 Examples of use in training and other similar environments

Of most relevance is the work of Schvaneveldt et al* who used the Pathfinder methodology to assess the change in knowledge structures of trainee F-16 pilots (n=60) after a week of training. They also compared the structures of the six most experienced pilots with a group of eight, least experienced pilots. They found that Pathfinder could produce intuitively appropriate solutions and that the change in network solutions as a result of the training process was reflected more in the least experienced pilots. While there was a tendency for individual pilot networks to become increasingly similar to each other over the course of training, they were not able to determine whether this was due to a common team experience or common training context and environment. It should also be noted that

the concepts that were used did not cover combat functional areas such course of action assessment or situation assessment.

In another related study, Schvaneveldt, Durso, Goldsmith, Breen, Cook, Tucker and de Maio (1985) showed that Pathfinder solutions could allow individual novice or expert networks to be discriminated using a pattern recognition algorithm. Although, it should be noted that a standard multi-dimensional scaling solution was equally effective in producing an accurate discrimination. Further, classifications using Pathfinder networks were found to be more accurate than classifications based upon the ratings alone.

Overall, these studies can be taken as providing a starting model for conducting similar studies with the Canadian Air Force, whether in a DMT or regular training environment.

In a non-military context, Goldsmith and Johnson (1990) have examined the degree to which general classroom learning can be represented by network models. The study population was undergraduate students in a course on psychological research methods. Thirty concepts were generated by a pool of faculty SMEs and were rated for similarity by students on a seven point scale in the first, eighth and fifteenth week of the semester. Student performance in the course was measured on three examinations and two papers. The results showed that the resulting PFNET solutions showed the type of changes over the course of the semester that would be expected as a result of student learning and became increasingly similar to the networks of the SMEs. Further, the similarity between an individual student net and that of the instructor was a good predictor of how much a student has learned about the specific domain of study, as assessed by course performance.

6.5.5 Limitations and issues not fully resolved

We have not been able to find any studies that have systematically assessed the validity or reliability of PFNET solutions in comparison with other MDS approaches. The face validity of Pathfinder solutions has been demonstrated in several studies (as cited in Durso and Coggins, 1990). Villachica (1999) has examined the discriminability of PFNET structure and coherence scores and found them to be high and statistically significant. The test-re-test reliability of relatedness scores was found to meet current recommendations (.8), but similarity and coherence scores did not.

Further, the relationship between knowledge structures as represented by PFNETs and relevant task performance has received little attention. One exception is the work of Goldsmith and Johnson (1990) whose showed that PFNETs predicted undergraduate course performance. Clearly, a parallel study to this needs to be conducted in a military training context, since the ultimate goal of such training is not simply to change knowledge, but to do it in a way that impacts positively upon the areas of operational performance under training. Notwithstanding such findings, there are a number of critical issues that the intended user of the Pathfinder approach must consider before implementing a study. These are outlined below.

6.5.5.1 Selection of concepts

Given the lack of prescription for a standardised set of procedures for developing the concepts for the domain of interest, the problem of how to generate appropriate, valid and highly germane needs to be addressed at an early stage in the study. An unsystematic and undisciplined approach will yield a set of concepts that are likely to produce network solutions that are uninterpretable or lack validity for the domain. To borrow a programming expression - "garbage in - garbage out"! Efforts should be made to obtain the relevant SMEs in sufficient numbers in order to provide the appropriate level of knowledge representation for the domain. In the case of Canadian Airforce training, decisions will

need to be made about whether these SMEs should be trainers, experienced pilots or possibly curriculum designers. Wherever possible concepts derived from training analysis should provide a good starting point for concept selection, such as the Mission Essential Competencies that have been used for this purpose in the United States. In Canada we could consider using the CF-18 Training Task List for the CF-18 Advanced Distributed Combat Training System.

Based upon what has been learned from the literature and from discussions with Pathfinder SMEs at Mesa, we recommend that concept development be an iterative process that has at least two cycles or refinement. Preferably, after the initial selection and reduction, a second sample of SMEs should be used to refine the set. A combination of intuitive, heuristic and statistical approaches may be the optimum way for ensuring the required level of validity and applicability of the concept set.

6.5.5.2 Choice of q

We have come across no systematic guidelines for, or validation of, procedures to select this value which essentially determines the density of the network. In some cases, the value has been varied to provide a solution that seems intuitively correct, in other cases it has been manipulated to produce equal density graphs across individuals. Further, it may be used to produce greater similarity between two sets of Pathfinder solutions, either when comparing different groups (novices, experts) or the same group at different points in time (before and after training). We have not seen any interpretations of how variations in network density across individuals or groups (when q is not manipulated) relate to properties of the underlying knowledge structure.

6.5.5.3 Sample size

The range of sample sizes found in published papers ranges from the teens to the several hundred. In some papers, subsets of the initial sample as small as six have been used to group data and compare individuals. In planned studies, sample size becomes an important variable when there is likely to be a high initial variance among factors of influence for the domain of interest. Thus, there would likely be a high initial variance in mathematical, logical and analytical skills in the Goldsmith and Johnson study that used junior level undergraduates. Such high variance could be expected to impact upon the variability of the underlying knowledge structures, thereby necessitating a reasonably large n to produce a representative and generalisable network solution. On the other hand, if we consider a sample of Canadian pilots who are starting CF18 training we would expect less individual variance in factors that are likely to affect their emerging knowledge structures. For example, prior training will have reduced the variance and resulted in common standards in areas such as basic piloting, navigation, planning and communication. Therefore, it may be possible to develop representative and reliable network solutions with samples in the 10-30 range. However, this remains a potential problem that may only be resolved when a study of this population is first conducted; as such, in order to minimise problems associated with high variance, loss of data due to drop out and outliers that have to be eliminated from the data set, we recommend that as large a sample as is practically available be sought, comprising at a minimum one full cohort of trainees, and preferably two.

6.5.5.4 Administration of the ranking procedure

There is a wide variation in the practices that are used for administering the pairwise rankings. Some studies report no instructions to guide participants, others have very detailed protocols. Some studies use a five point scale, others a seven point. In order to minimise inter subject variance and to maximise motivation we recommend that a standard protocol be adopted for test administration. This would contain the following elements: initial briefing by respected military SME to provide context

and stress importance of the study; some form of extrinsic incentive to enhance motivation and specific guidelines to the participants on the use of the scale. With respect to the latter we recommend adopting the approach of Goldsmith and Johnson (1990) who encouraged participants to be diligent in using the full range of the scale and to use ratings at the end of the scale (1,2,6,7) when they were more certain of the meaning of the concepts, and intermediate ratings when less certain. They were also guided to make quick intuitive judgments rather than an elaborate analysis of the relationship among the concept pairs.

6.5.5.5 Interpretation of links

In Pathfinder analysis links in a PFNET graph are not labelled in any way nor are they differentiated from each other nor do they indicate anything specific about the relationship among the nodes that they connect, but rather represent a general association between the concepts represented by the linked nodes. This gives rise to several questions in interpretation. For example two concepts with the same links in a novice or expert PFNET do not imply that the novice has the same understanding of the relationship of the concepts. Thus, the link in itself is neutral with respect to the issue of how concepts are related.

In an attempt to overcome these problems and in the absence of an existing methodology for labelling links, Cooke (1990) has proposed and evaluated a technique for this purpose to permit improved interpretation of PFNET solutions. Her solution involves two steps: link classification using a sorting of link labels generated by an SME on a pairwise basis together with a clustering analysis of the results, followed by cluster rating and labelling, in which relationships are attempted to be defined for meaningful clusters of links. While this approach remains promising, it should be noted that the links in the study represented fairly simple and obvious relationships and would not be appropriate for adoption into a flight training context. Further, there are issues concerning subjectivity in naming links and issues of high variance among expert raters. Notwithstanding such limitations, we recommend that future studies that incorporate Pathfinder approaches to assess changes in knowledge structures in the aircrew training domain, consider adoption of procedures that would allow interpretation of the meaning of the structural *relationships* among concepts as well as those comparison techniques discussed above that allow for comparison of the overall structures themselves.

6.5.5.6 Analytical techniques for evaluating and comparing PFNET solutions

The literature reveals no standardised approach as to how to make formal comparisons among different PFNET solutions, whether comparing individual to individual, individual to group or group to group. The coherence of the network is a popular choice for examining expert-novice differences, whether at the group or individual level. Goldsmith and Johnson (1990) have evaluated four comparison techniques -Pearson correlations between two sets of raw proximity ratings, correlations on MDS distances, correlations on graph theoretic distances in PFNETS and a set theoretic comparison of PFNETS that generates a metric C. For the MDS solution, Euclidean distances are calculated between all pairs of points in n dimensional space and then the correlation between corresponding distances in the comparison networks is calculated. The set theoretic approach (see Goldsmith and Davenport, 1990) a single quantitative metric of closeness C, that can vary from 0-1. Thus, C represents a preferred approach for comparing the configuration or structural similarity across different Pathfinder networks. In reviewing the changes in knowledge structure over time, the authors reported that the MDS solution appeared to be the most sensitive. However, in correlating knowledge structure against that of instructors at the end of the semester the metric C produced the highest correlations (.74) of the four comparison methods, although all reached statistically significant levels. Further, when correlations of the other measures are held constant, the partial correlation coefficent of

the C metric still correlates significantly with performance. They conclude that PFNETs provide a better basis for comparison than the raw proximity ratings, and the configural comparison among different PFNETs using C is a better choice for assessing network similarity than MDS correlations. This implies that the C metric better reflects structural similarities in the comparison nets that are important representations of underlying knowledge.

6.5.6 Application to Distributed Mission Training

The most obvious application of the Pathfinder approach would be to conduct a similar study to that of Schvaneveldt et al (*) using CF18 trainees in a Canadian context. That is, the knowledge structures of a cohort (or more) of trainee pilots would be assessed prior to the CF18 course and then assessed at the end of the course. The change in structures as represented by PFNETS over time would be analysed and compared with expert PFNETs. Further, the PFNETs of more and less experienced pilots within the group could be compared. To enhance the validity and value of the study it would be useful to also collect performance data for the trial participants. Such data could come from ratings from trainers, course examination results or objective measures of operational performance in a simulator. The outcome from such a study would be of benefit to both the scientific and operational community. For the former, the Pathfinder methodology can be assessed first hand and various analytical procedures compared. It may also be possible to determine the degree to which structural knowledge (compared with skills and psycho-motor competencies) actually contributes to performance effectiveness. It may also provide a basis for analysing the influence of learning (training) at the skill, rule or knowledge levels. A further interesting analysis might involve the degree to which *team* structural knowledge is influenced by the DMT environment and whether the team mental models extend beyond the specific operational grouping formed by separate participating countries. For the operational community it may be possible to identify which clusters of concepts are most influenced by training, which clusters might be taught together, which aspects of the network correlate with operational performance and whether individual student capability and learning can be assessed from PFNET solution.

Finally, details are provided in an Appendix of the complete logistical requirements to conduct a Pathfinder trial in an AF training environment.

6.6 Situation awareness: concepts and measurement

In this section we provide a brief overview of the concept of SA and its usage in the AF community. The goal will be to define SA in an operational manner that allows it to be assessed in a flying mission context and to review and recommend measures and methods for its assessment.

6.6.1 Definitions

Before proceeding with a more extensive discussion of definitions, we provide below some selected (not comprehensive) examples of how situation awareness has been defined in the literature.

> Endsley (1988): *"the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future"*.
>
> Whitaker and Klein (1988): *"the pilot's knowledge of his surroundings in the light of his current goals"*.

Sarter and Woods (1991): *"situation awareness is the accessibility of a comprehensive and coherent situation representation which is continuously being updated in accordance with the results of recurrent situation assessments"*

Smith and Hancock (1994): *"situation awareness is up-to-the-minute comprehension of task relevant information that enables appropriate decision making under stress", and "adaptive, externally-directed consciousness".*

Given the diversity of these definitions, perhaps an obvious starting point is to look at what is understood about each of the two terms that together form the concept of situation awareness.

Pew (1994, 1995), working from an aviation perspective, defines a **situation** as:

"..a set of environmental conditions and system states with which the participant is interacting that can be characterized uniquely by a set of information, knowledge and response options".

Note that one characteristic that Pew identifies separately from the generic environment is the state of a system. This seems somewhat redundant given that a system is already one particular aspect of the environmental conditions. However, it does serve to remind us that when we are referring to a situation we include both the environmental component and all relevant human-machine systems within it. The second important element of Pew's definition is the notion of **interacting** with the environment, which implies active rather than passive behaviour.

Turning to the second word of the concept, Pew defines the elements of **awareness**, given a situation as follows:

- The current state of all relevant variables of the system
- The predicted state in the near future
- The information and knowledge required to support current activities
- An activity phase
- A prioritised list of current goals (and sub-goals and time)
- The information and knowledge needed to support anticipated near future contexts

It is important to note that this definition clearly extends the concept beyond simple awareness of just the *spatial* relationships in the environment. There are three additional core concepts that expand the definition, which we illustrate below with examples from the aviation domain.

First, awareness includes knowledge of the *state* of the environment. In an AF mission context context, the state of the environment would include factors such as: the status of weather and terrain and the location and disposition of relevant enemy and friendly units.

Second, awareness encompasses knowledge of the individual's own state and goals. This would be exemplified by a pilot's awareness of the flight/navigation plan, the sub-goals along the way, the sequential order and list of tasks that must executed to achieve the plan and an awareness of the overall mission state.

Third, awareness involves an appreciation of how the future situation will change, accompanied by a comprehension of what new information and knowledge will be required for that future state. In an AF mission, new information will arrive en route that will affect the need to change some aspect of the plan at some future time, tactical circumstances will change as a result of successful or unsuccessful engagements along the way and the pilot must have awareness of what new information needs to be obtained as a result of anticipating future changes in the situation.

Pew and Sarter and Woods (1991) outline multiple elements of awareness that need to be considered in defining the concept.

Spatial awareness: an appreciation of the spatial location of ownship, others in the formation, other relevant participating units, enemy formations and terrain.

Identity awareness: the pilot's awareness of key navigational points and decision windows..

Temporal awareness: knowledge of the changing spatial picture over time. Projecting flight vectors and trajectories. Temporal awareness is a key component for performing the mission in accordance with the plan and knowing when critical timing elements are jeopardised..

Mission/goal awareness: At the highest it may be the tactical mission objectives for that unit (e.g. ingress, deliver ordinance, egress, deal with threats).

System awareness: attention to the relevant information from aircraft systems (e.g. flight dynamics, weapons, navigation, sensor).

Resource and crew awareness: in a DMT environment this would involve the monitoring of other distal participating units and a comprehension of how changes in their status and capability impact upon ownship/formation ability to meet mission goals.

6.6.2 The work of Endsley

In any consideration of the concept of situation awareness the ongoing contributions of Endsley have been central to the emergence and acceptance of the concept of situation awareness within the Human Factors community.

Endsley's basic, initial definition of situation awareness shares much in common with Pew by describing situation awareness as

> "*a person's state of knowledge or mental model of the situation around them*". This definition is further expanded as follows: "*the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future*" (Endsley, 1988).

This latter definition probably remains the most quoted and universally accepted in the field. Not included in this expanded definition, but core to Endsley's concept of situation awareness, is the notion of *integrating* disparate pieces of information from the environment and relating them to the task in hand or the current goals.

It should be noted that there is a subtle change in emphasis between the above two definitions of situation awareness, with the first describing a "state" and the second describing processes. This distinction has given rise to some controversies, which we shall visit later. Somewhat contradictory to her own definition that includes *processes* of perception and comprehension, Endsley insists that "situation awareness is a *state* of knowledge about a dynamic environment" (1995b). This is to be distinguished from the processes that give rise to the knowledge. In a more recent definition of situation awareness Endsley (1996) states that it is a person's mental model of the current state of a dynamic environment.

Endsley provides a clear description of the underlying processes that are involved with the creation and maintenance of situation awareness using terms and language derived from information processing models of human performance. These major cognitive components involve working and long-term memory, attention, and mental models or schema.

In Endsley's (1994 and 1995 a,b) conceptions, situation awareness is to be clearly differentiated from both decision making and action (or performance). Situation awareness serves as the cognitive *antecedent* to the decision making process by providing the appropriate information for decision input. Situation awareness is part of the individual's internal model of the environment. It follows, therefore, that if situation awareness is degraded less than optimum decision making will occur. The corollary is not true, however, having good situation awareness is no guarantee that the appropriate or correct decision will be made. This may occur because an individual has limited experience in making decisions, poor decision tactics and inadequate decision choices.

Endsley, like Pew, has characterised three fundamental aspects of situation awareness, which she refers to as "levels" as follows.

Level 1: Perception of the elements of the environment
This process requires that the individual engage in the processes of search, detection, recognition and identification of the relevant status, features and attributes of the environment that are pertinent to the goals in hand. In an AF mission context this would involve the standard in-cockpit, instrument scan, the monitoring of aircraft systems, the search and detection of external objects of interest on the ground or in the air.

Level 2: Comprehension of the current situation
Comprehension is achieved through the integration and synthesis of the relevant, disparate information that is acquired through level 1. In this process the pilot goes beyond the simple knowledge that the information is present in the situation and seeks to determine the significance and meaning of it. In some cases, this may produce a holistic picture of the environment. In complex environments the ability to achieve this level of comprehension will be dependent upon the skill and knowledge that is acquired through experience. Pilots refer to this as "having the picture" or, when SA level2 breaks down, "losing the picture".

Level 3: Projection of future status
This refers to the ability of an individual to anticipate or envisage the future status or actions of the elements in the environment over a near time frame. This is a critical component of SA in any AF mission context. The pilot must be able to continuously switch between monitoring the current state to anticipation of the future situation. The time frame for this may be short, i.e. in a merge with a hostile aircraft and anticipating its tactics, or long, for example estimating implications for completing the mission goals when there has been a significant deviation from the planned route, timing, resources available or enemy disposition.

6.6.3 Suggested operational definition of SA

Like Endsley, we believe that SA is the key to good decision making in a AF mission environment. Decisions can only be as good as the information on which they are based - particularly for highly trained and practised tactical decision makers such as pilots. Thus, while it would be informative to study decision errors by pilots, we would learn little about the underlying reasons that may have resulted in the faulty decision.

If the basis for good decision making is SA, we may wish to consider what is the best way to describe and define this concept in a mission relevant manner that will allow it to be studied and quantified in a reliable and valid manner.

While, we can base the definition on the above examples taken from the scientific literature, we believe that the definition should also take cognisance of the way SA is thought about and talked about in the AF community. Our experience and conversations with AF SMEs indicate that there are two

major distinction in common usage. *Local SA*: this typically means the pilots integrated tactical picture of the immediate and dynamic short-term situation. It is most often used in the context of an engagement with an enemy unit. When a pilot says that he has lost the picture, it means that he is no longer able to integrate all of the disparate bits of information into a coherent meaningful pattern. *Global SA*, on the other hand, appears to be more associated with the larger mission context. Thus, after an engagement with an enemy aircraft, pilot's note that their first task is to regain the global picture. This will not only involve re-integration with the local formation but developing an understanding about how mission goals (and tasks to achieve the goals) may have been changed (or not) while the pilot was locally pre-occupied. Regaining the global picture will require a proactive seeking of information from a variety of sources and will usually be accompanied with frequent communications.

We can integrate the definitions from the literature with this AF domain usage of the concept to arrive at the components of SA that could direct the process of measurement as shown in Table *. Spatial, temporal and system awareness are the major components that support Local SA, particularly in the context of an engagement. Spatial awareness largely involves integrating spatially separate elements of the tactical situation into a coherent picture and predicting their future relationship. These elements include ownship speed and location, other nearby participating units in the formation, terrain and the disposition and movement of enemy units. The problem for the pilot is largely one of integration of information and projection of future status. The only element of level 1 SA would be the task of visually detecting the tactically relevant unit, for example searching the sky for the exact location of a hostile that has already been detected and located by sensors or radar. Temporal awareness will largely involve levels 2 and 3 SA, as the pilot attempts to estimate closing and intersecting trajectories and extrapolate future positions in an engagement. System awareness would require some level 1 SA as the pilot must detect relevant information provided by aircraft systems, however, since this is such a highly trained skill it may only involve detection of deviations or anomalies. These could take the form of critical aircraft system failures or problems, sensor alerts or urgent communications. The major task for the pilot however is to continuously integrate the generic system information to ensure that the aircraft performs as planned for the circumstances at hand. Much of this skill is again highly overlearned and practised and may require less direct attentional processing than maintaining spatial and temporal awareness.

Type of SA	Components	Level
Local	Spatial awareness	Some 1, but largely 2 and 3
	Temporal awareness	Largely 2 and 3
	System awareness	Some 1, but largely 2 and 3
Global	Identity awareness	1 and 2
	Resource and crew awareness	1 and 2
	Mission awareness	2 and 3

Table 9: SA concepts for potential measurement

Global awareness is made up of the components of identity, resource/crew and mission awareness. Identity awareness requires the ongoing monitoring of position with respect to the navigation plan and mission timeline, and also the timing for the performance of critical tasks. Hence, this is largely an information detection and integration process. Resource/crew awareness largely involve the detection of information concerning critical changes to the current state of crew or resources and the comprehension of how such disparate information may affect the mission plan. Mission awareness is

accomplished through Level 2 and 3 SA in that the pilot is involved in the integration of disparate data to form the mission picture. Such data include the mission plan, data resulting from identity awareness and crew/resource awareness, location of other friendly units, and the location of hostile and other significant elements to be encountered further in the passage to the mission goal. Since there will be dynamic changes to these elements en-route, the pilot must understand the implications of such changes on the future mission picture, i.e. level 3 SA. Thus, mission awareness is highly dependent on information that is provided on an ongoing basis by the monitoring of identity and crew/resource data.

With these sub-components of SA defined in AF mission context, we can now turn to the task of developing MOPS that are designed to assess the different facets of SA as represented in the above classification scheme. As a background to our recommendations for the types of MOPs that may be most applicable, we provide a brief review of the literature on measurement issues in SA.

6.6.4 The Measurement of SA

In this section we address the difficult and somewhat controversial issue of the measurement of situation awareness. The major topics covered are *what* to measure and *how* to measure. For the most part, we deal with the measurement of situation awareness in the general case, since most of the arguments and conclusions will also apply to the DMT context. Where feasible we provide aviation exemplars to clarify specific points.

6.6.4.1 What to measure

If we regard situation awareness as being a complete abstraction that only serves as an operationally useful concept, then it would follow that no means of measuring or quantifying it would be possible. Alternately, if the abstraction represents simply a new label for a process in which an individual recombines or extracts information using standard information processing mechanisms (e.g. attention, working memory), then it follows that the focus of measurement might be on those underlying mechanisms and processes. Of course, prior to the emergence of the concept of situation awareness, that was exactly the focus of research. This approach is favoured by Sarter and Woods (1995) who view situation awareness as a convenient label for an aggregation of such processes and argue that an understanding of situation awareness is best served by measuring and understanding the underlying processes.

If the level of research interest or explanation is in these underlying processes then measures of spatial, attention, memory, perceptual and cognitive abilities and processes will be the appropriate level of focus. In some instances this may provide useful information about how pilots allocate attention, what information takes priority and how novice and experts differ. To illustrate and clarify the relationship among the various levels and types of measurement and associated human processes, we have adapted Endsley's (1996) process model and applied it to the DMT context as shown in Figure 3. Some of the specific measures of situation awareness shown in the chart will be explained later in this section.

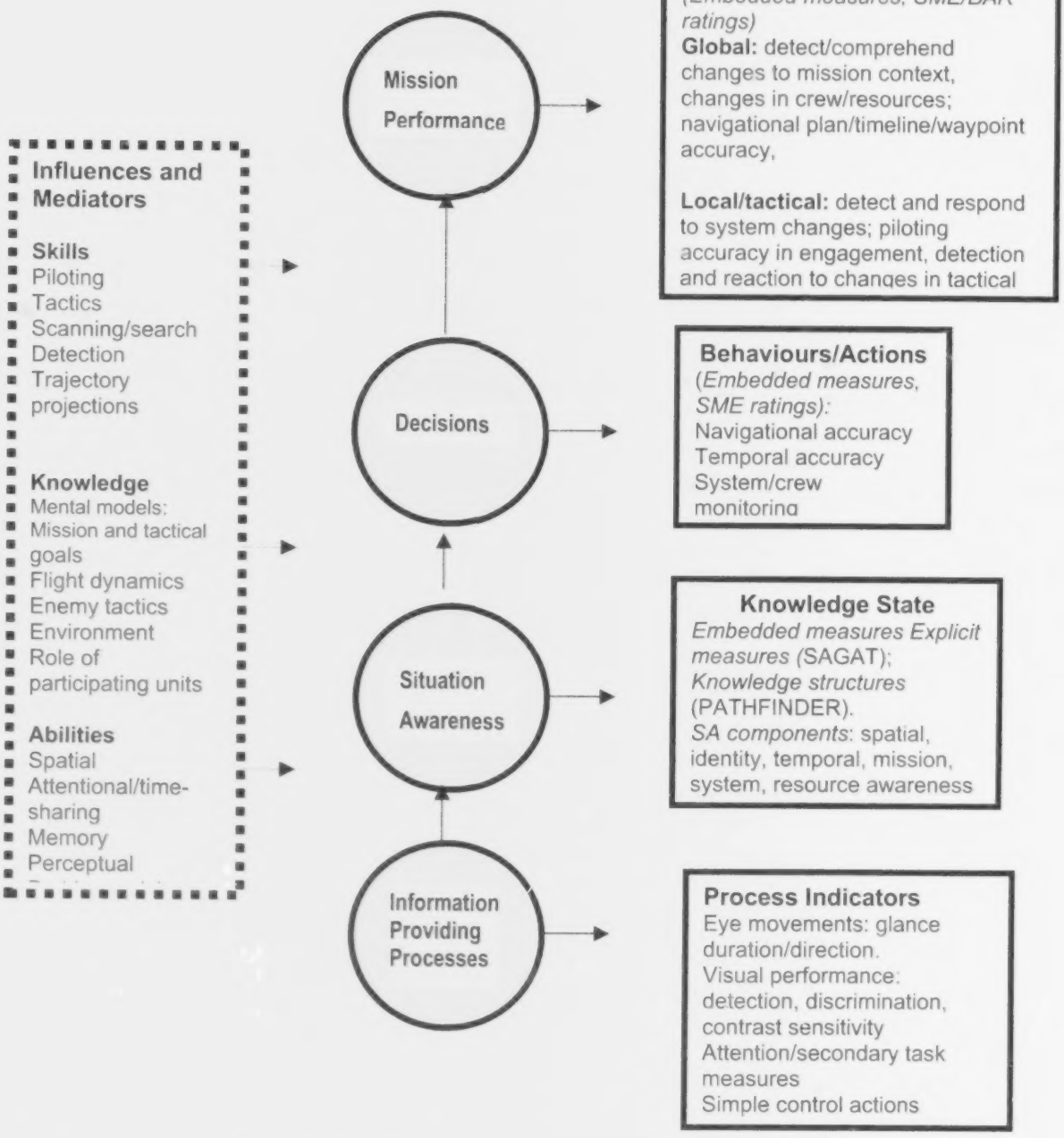

Figure 3: Situation Awareness process and Measurement Model
(adapted from Endsley (1996))

With this model in mind, the concept of situation awareness provides a means of re-framing some of the measurement questions in a new light. In particular, we can ask questions about the *process* of how an individual acquires and builds situation awareness, the *content* and relationships of the information currently held that contributes to situation awareness, and how the individual plans and anticipates future information needs. Thus, we take the perspective that while the abstract concept of situation awareness is inherently unmeasurable, it does provide a useful framework for asking questions that have the potential to be addressed with appropriate measurement approaches that yield metrics and data of value.

As we have seen from the previous discussion, situation awareness has been invoked for the most part with respect to complex domains of behaviour and task environments. The following list represents some generic characteristics of these application domains.

- Moderate to high information rates
- The requirement to perform multiple tasks
- The need to search and extract information from many and usually complex sources of information
- The need to consider and plan for future actions as well as monitoring current tasks
- Behaviour serves several different goals.

Under such circumstances, the information content of an individual's situation awareness will be correspondingly complex and changing. Thus, a major problem for the measurement of situation awareness is defining what is the relevant *content* of situation awareness that should be measured in a particular context.

This problem is underlined by Sarter and Woods (1991) who note that

> *"Attempts to define the critical contents or components of situation awareness in general suffer from the fact that, given the dynamic environment....., the relevance of data and events depend on their context"*.

Thus, similar data under different circumstances may or may not be necessary to support the user's situation awareness depending upon the current goals and task at hand.

This does not necessarily mean that situation awareness is inherently unmeasurable, but that it is a complex task to define what the reference, benchmark content should be. For example, in the DMT context, the assessment of whether a pilot has appropriate situation awareness may not be able to be simply measured using a constant set of variables across all of the situations of interest. Instead, each situation must be separately assessed in terms of the pilot's current goals with a view to identifying the specific information that a pilot needs to maintain in situation awareness to support those goals at a particular point in time. This requirement is stressed by Smith and Hancock (1995) who note that:

> *"Until and external goal and criteria for achieving it are specified, examination of greater or lesser degrees of situation awareness or even loss of situation awareness remains impossible"*.

In general we can say that for a person to perform effectively in dynamic, multi-tasking environments, it follows that she/he must sample widely and frequently among the information sources to acquire and maintain the necessary level of situation awareness. Therefore, the scope of the measurement (Vidulich, 1995) needs to be correspondingly broad and multi-dimensional. A situation awareness metric with wide scope has the potential to sample broadly among the identified components relevant to the context of the behaviour in question. In contrast, a measure with narrow scope fails to capture

the required behavioural complexity. For example, in the case of DMT, a narrow metric of situation awareness would be the pilot's assessment of current location with respect to the mission plan, or his tactical placement in combat.

The requirement for a multi-dimensional approach therefore rules out using global measures of effectiveness to provide a valid estimate of the broad spectrum of factors that influence current situation awareness. In a DMT context accuracy of ordinance delivery and the number of enemy destroyed would be example of a global measure of performance. In general, for complex systems, we have argued that global measures of effectiveness (Matthews, Webb, McCann, 1997) are insufficiently diagnostic of situation awareness. Therefore specific measures need to be developed to measure the processes themselves that comprise SA and hence lead to overall performance effectiveness.

Finally, the conceptual model that we provide provides a means for identifying the overall factors that are likely to influence situation awareness. For any particular context, the current goals of the pilot can be evaluated against the model, and the important variables that are present to influence situation awareness provide a means of identifying what aspects of situation awareness should be measured.

6.6.4.2 How to measure SA

The major approaches to measuring situation awareness are categorised below. In each case we provide a general overview of the approach, some selected examples drawn from the literature and finally a brief discussion of the merits and disadvantages of each approach, particularly with reference to their applicability to the DMT context. The specific measurement approaches discussed are:

- Global/outcome measures
- Task interruption/probe techniques
- Implicit measures
- Subjective ratings: by participants or trained observers
- physiological measures.

In evaluating candidate measures for their suitability for assessing the various dimensions of situation awareness, the following criteria should be considered: validity, reliability, diagnosticity and practicality.

Validity There are three separate aspects of validity. *Content* validity is normally considered to an attribute of a measure that is related to how well it represents the behaviour being sampled. This is a very important consideration in the measurement of situation awareness, since its content and internal relationships are complex, then this should be reflected in the measures used. Thus, using a dual task paradigm to assess the spare attentional capacity as a measure of situation awareness would have low content validity. To achieve content validity, we must first identify the full *range* of information that would be expected to be part of situation awareness for a specific context, and then ensure that we develop appropriate measures to sample this information only (i.e. no extraneous or irrelevant domains of information).

Criterion validity refers to the extent to which the measure of choice will actually correlate with some independent and objective measure of the behaviour in question. This may often be difficult to establish because no such independent metric exists, or because the individual's experience with a task will be a confounding variable.

Construct validity refers to the extent to which the metric actually reflects the hypothetical, psychological construct that it is supposed to measure. Clearly, situation awareness is such a construct. A complex procedure is required to establish construct validity. First, we need to show which behaviours are specifically the result of the construct and not related to other constructs. Second, we identify related constructs that are logically related to the construct of interest and identify their associated behaviours. The goal is then to demonstrate that the behaviours related to the construct are different and uncorrellated with behaviours unique to related constructs. Unfortunately, such procedures often difficult to do and their outcome may be ambiguous. A failure could either mean that the measure had low validity, or that the underlying theory is incorrect. This circularity and interdependence of the measure and the construct present a continuing challenge to the development of both theory and measurement, particularly in the case of situation awareness, which is shown to be conceptually related to other information processing constructs such as working memory, attention and long term memory.

In general, in our assessment of measurement approaches, we will evaluate the validity of a measure in terms of its content validity primarily, unless there are empirical data to available to inform assessment of criterion or construct validity.

Reliability concerns the degree to which the measure will produce similar data given repeated tests under different conditions. Reliability can refer to the consistency of the measure over different time periods, different experimental conditions and across individual subjects.

Diagnosticity refers to the degree to which a measure informs us about the process in question and is related to aspects of criterion and construct validity.

Practicality in the present context concerns the ease with which the measure can be implemented in a DMT or other mission test environment.

At the end of this section, we provide a rating for each candidate measure for each of these dimensions using a simple rating scale of low, moderate and high. These assessments are based upon a combination of our knowledge and interpretation of the available literature and our experience in using these measures in a variety of application contexts.

Global/outcome measures

These can be thought of measures that are related to the typical behavioural outcomes required for performing a task. In a DMT context, potential global measures might be number of enemy destroyed, percent of armaments hitting target, kill/loss ratios. Global measures thus represent an output behaviour that is usually the observable end-product of a number of underlying and unobservable processes. For the most part, therefore, global measures lack sufficient diagnosticity and specificity to usefully inform us about the processes of acquiring and maintaining situation awareness. A further confounding element comes from potential interactions with other external variables that influence the behaviour outcome.

Task interruption-probe techniques

This category comprises a number of approaches that may use probes or questionnaires, however, common to both approaches the ongoing task flow is interrupted in order to administer the measure. During the interruption, displayed or available information may be frozen, but, more typically, information is removed from the individual's view. In most studies this has involved blanking a display screen. The interruptions are normally scheduled at random intervals to prevent anticipation.

The basic assumption of this approach is to tap into the individual's perception and knowledge of the current situation. The resulting data may be able to be compared against "ground truth", that is, the actual situation specific information at the time of the interruption. Hence the usual measure of performance is percent items correctly detected, recognised or reported. Examples of such techniques are shown in the box labelled "Knowledge State" in Figure 1.

This technique has been widely used in the aviation domain to evaluate different display options (Marshak, Kuperman, Ramsey and Wilson, 1987), and by Fracker (1990) to examine attention in situation awareness. Support for the construct validity of explicit measures using freeze probes for hazard detection has been provided in an aviation context by Olmos, Liang and Wickens (1995).

In general, this methodology involves an evaluation of a limited, selected subset of the total information that may be currently held in situation awareness. To overcome this limitation, Endsley has developed a variation of the methodology called the Situation Awareness Global Assessment Technique (SAGAT), which is designed to measure situation awareness comprehensively at all three levels. The SAGAT method has been widely adopted with reports of success in diverse environments, for example nuclear power plant control rooms (Collier and Folles, 1995) and driving simulation (Gugerty and Tirre, 1999).

The basic process in the developing the methodology involves the following steps:

- determining the specific knowledge that is expected to be part of an individual's situation awareness at a particular point in time
- considering which level of situation awareness should be the focus of the probes
- developing probe questions that will tap into that knowledge
- developing appropriate response methods
- determining a schedule and duration for the task interruptions

The general assumption underlying the methodology is that at the time of the interruption, data are available to the individual in working memory and that these are amenable to the probe questions. Because the SAGAT probe battery is administered immediately after the removal of information, the content of situation awareness may be probed before working memory completely decays.

Since this is a core assumption of the methodology, its validity has been evaluated by Endsley and others. The claim is made (Endsley, 1995c) that the information content of situation awareness may be available as much as 5 or 6 minutes after the freeze. What remains to be addressed in the literature, however, is the generality of this finding, and whether it pertains to all of the current information contained within situation awareness. To elaborate, the contents of situation awareness may comprise (i) data recently sampled from the environment and held in short term, memory, (ii) a mental picture of the current situation that derives from previously integrated data sampled from the environment incorporated into a knowledge structure from long term memory and (iii) a low level appreciation of the ongoing state of semi-automated or automated processes. None of the literature reviewed has provided an answer as to whether all of these types of information are equally amenable to a probe technique and whether the measure remains stable and reliable for the length of time that Endsley claims.

A further problem for this approach in the context of DMT is the highly dynamic nature of SA, particularly during times of tactical engagement. The timing of a probe may be critical down to a second or two and the nature of the information processed by the pilot may be so transitory that it is no longer available for retrieval from STM. It may be possible to obtain some crude data from such an approach concerning a pilot's comprehension of the current location of a threat or its likely trajectory

over the next few seconds. However, it would not appear to be sensitive to assessing factors that contribute to loss of SA or the complex interrelationships between time, space, ownship, enemy ship and terrain.

An additional concern arising from the methodology is how the interruptions interfere with, intrude upon, or contaminate, the ongoing task at hand. One aspect of this problem is how the individual subsequently performs on the task following the interruptions. Endsley has attempted to evaluate this in a simulated air-combat context, by looking at performance in trials with and without probes (Endsley, 1995c). Her results suggest no difference between the conditions in terms of number of "kills" or "losses", nor was there an effect of the number or duration of freeze interruptions. However, the data rates in her study may be significantly lower than in a typical combat mission.

The generality of this claim that there is no effect of the interruption on task performance and associated situation awareness must be questioned. First, the global measure of performance used by Endsley (while no doubt practical in combat terms) may be insensitive to actual changes in workload or situation awareness induced by the interruption as the pilots try to regain the picture after the interruption. More importantly, however, the results may not generalise well to other situations in which the individual has a greater level of multi-tasking to manage. Hence, the interruption may cause a major disruption in not just one task (e.g. flying the aircraft in the Endsley study), but all other ongoing tasks. This may then result in an inability to pick up the pieces on all tasks to arrive at the same level of global situation awareness that existed before the freeze interruptions. Such problems have in fact been reported by Brickman et al (*) who found that freeze interruptions were a source of annoyance and resulted in a loss of performance in high information rate, multi-tasking environments. This has been confirmed by (Small 1995) who reported that the interruptions presented unnatural problems for individuals in terms of regaining situation awareness in dynamic environments.

Endsley (1995c) provides some clear guidelines on how to implement the SAGAT methodology. There should be sufficient ***training*** to familiarise subjects with the probe questions; three to five trials are recommended. This training should follow the actual test procedures by interrupting the main task but giving subjects more time to respond than they would in the actual experiment. Second, if the paradigm requires the simultaneous collection of data using performance measures, then a second condition should be run without the SAGAT interruptions, to check for any possible cross contamination of performance data by the SAGAT methodology. (Note: such an option may not be logistically possible in the context of scarce military personnel resources). In terms of ***procedures***, subjects should be instructed to perform tasks as they normally would and not to use different strategies (see commentary below). During the period of the freeze, subjects should not have access to information from displays or other relevant information sources.

The more difficult choices in the methodology involve the actual choice of ***probe queries*** on any trial. If we are attempting to sample all three levels of situation awareness in a complex domain, a choice must be made of which particular information to select as the source of a query on any trial. This is further constrained by the need to keep the freeze interval sufficiently short to minimise disruption to the primary task and to reduce the mental workload in re-gaining situation after the freeze interruption. Intervals in the range of 1-2 minutes are frequently found in the literature, but Endsley (1995c) reports using an interval as long as 5 minutes with no adverse effects. We suggest that the more dynamic the environment and the greater the number of tasks that must be simultaneously performed, the shorter the probe interval should be. In general, the best approach will be to sample randomly among the relevant potential probes that are potentially available at each trial. In particular, measures need to be taken to ensure that the probe sample is sufficiently diverse that subjects cannot learn to anticipate the likely probe content as the experiment progresses. This would result in an overestimation of

performance in one aspect of situation awareness, against another area that the subject learns is less likely to be sampled. In some cases, it may be necessary to impose some constraints over the random selection, if the prior segment of the trial did not actually contain content that related to a particular level of situation awareness.

A further consideration is ***when*** and ***how frequently*** to freeze the trial. A random inter-trial interview is preferred to prevent anticipation effects. Depending upon the complexity of the situation and how long it takes to build situation awareness, it may be necessary to wait several minutes or more before the first interruption. For similar reasons, trials should be spaced apart by at least one minute (Endsley) and probably longer, again depending upon the domain complexity. The vagaries of a purely random selection may result in unrepresentative sampling of the context if the environment complexity and events change significantly over time and few trials and subjects are available to normalise this over a larger sampling. In such cases, the experimenter may impose some selected or stratified sampling to ensure representativeness of the data, while maintaining sufficient randomisation to prevent anticipation and other biases.

The ***frequency*** of freeze interruptions will be dependent upon trial length and the number of data samples required. This can become a major challenge in implementing the SAGAT methodology in a manageable and practical way. In a complex environment, we may have five measures of different aspects of situation awareness at each of three levels. Assuming a sample of at least five trials (ten would be preferred) for each, this gives a potential of between seventy five and one hundred and fifty probe questions. If we assume a maximum of five probes per interruption (which may be too many, depending on task complexity), this would mean fifteen to thirty data samples. Clearly, this would result in a frequency of interruptions that would be far too disruptive for most tasks with a half-hour task length, or possible even an hour (one every two to four minutes). To accomplish the required number of data samples, in practical terms it will therefore require that the several sessions be conducted with sampling spread across these. This requirement will have the undesired side effect of making the data suspect to other contaminating variables from the subject or test procedures. They also transform radically the DMT context from a that of executing a mission in a smooth, flowing manner to one in which progress towards mission goals would be a series of discrete components that would challenge the content validity of the methodology. Thus, to avoid such possible contaminants it would appear that the frequency of conducting probes should be highly constrained to prevent fundamental changes in the way a mission is executed. This in turn means that only a limited sample of behaviours may be sampled and/or insufficient data are collected to provide the required reliability of measurement.

The ***scoring*** of responses to the test probes may require some expert judgment. Generally, this will have been established prior to the experiment, when domain SMEs provide their expert opinions on what information to sample in a particular context, what probe questions are appropriate to assess this information and what appropriate or correct responses will be. Sometimes the correct answer will always be explicit and unambiguous (e.g. "Did you see any evidence of Fishbed in our last scan?"), in other cases some SME judgment will be required.

A final consideration in the use of such probe techniques is where does the operator get the information in response to the probe query, working or long-term memory? The general assumption is that the information is retrieved from working memory, however, information from long-term memory may also form part of current situation awareness in complex situations in which the individual is highly experienced. Therefore some care is needed in the choice of probes and interpreting the responses, if one is trying to establish whether the information concerned reflects data sampled from

the current situation or contributed from long-term memory based upon stored information patterns accumulated with task experience.

Studies on the psychometric properties of SAGAT

The literature search revealed very few studies that have looked at questions concerning the reliability or validity of the SAGAT methodology. In one of the more comprehensive studies to examine this issue Fracker (1991a), in a simulated air combat context, produced a number of findings that question the psychometric robustness of the measure. First, reliability was low, with a test/re-test correlation of 0.6. This is a level that many researchers would consider unacceptably low. Even worse, the reliability of SAGAT probes for the item location was close to zero (note that location of objects in the visual field is of very high saliency in air combat, and one would expect to find such information within situation awareness). It is possible that Fracker obtained such low reliabilities by using a test/re-test approach, so that if individual subject's performance varies in an inconsistent manner from one test session to another, this will increase the overall error variance and reduce estimates of reliability. A better approach might have been to collect many more individuals pieces of data for each variable of interest and then to look at measures of reliability within a test session.

The low reliability of the location probe measures suggests potential problems for applying the SAGAT methodology to the driving environment, where situation awareness of the exact location of objects in the driving environment may be important for many tasks. It is possible that the problem arises in the context of the particular probes chosen by Fracker and the particular criteria he used to establish accuracy. Location probes using the SAGAT methodology have been tried successfully in a driving simulation context (Gugerty and Tirre, 1995a), although no specific psychometric measures were employed in this study to evaluate the reliability of the data obtained.

With respect to validity issues, Fracker (1991a) found reasonable levels of criterion validity (once corrections were performed for low reliability) for SAGAT location probes. Fracker also obtained some limited support in one experiment for construct validity, in that SAGAT probes, which tapped into verbal or spatial processes, generated data to support the notion that these are independent components of situation awareness. However, in general, Fracker was unable to conclude that there was sufficient evidence to support the overall construct validity of the measures of situation awareness in a way that would clearly differentiate them from measures of working memory.

Overall, we have yet to find comprehensive studies that have evaluated the soundness of the psychometric properties of SAGAT. There are two notable deficiencies in the studies we have reviewed. First, in most studies, experimenters tend to evaluate only a small subset of the information potentially contained within situation awareness, and hence use a limited number of measures. Second, few studies have evaluated measures that tap into how an individual comprehends the "big picture", or how various pieces of information within situation awareness are integrated together to form a comprehensive picture (i.e. Endsley's level 2).

6.6.4.3 Variations on the SAGAT approach

Where the task environment may be unconducive to the freeze-probe approach, it may be possible to first make a video tape recording of the task domain and then subsequently play this back for analysis. Either the original or naïve subjects could then view the playback with suitable probe interruptions introduced. This method has been reported to work successfully in the nuclear control domain (Sheehy, Davey, Fiegel and Guo, 1993). However, its utility for mission-related, pilotting tasks must be somewhat questionable, since the fundamental characteristic of that environment is the pilot is always completely in the loop. Also in flying a mission there are specific, real-time goals to be

achieved; these determine associated tasks to be sequenced, which in turn control and determine the content of the pilot's situation awareness. If the pilot is taken out of the loop and views the information passively, then there are no contextually relevant goals to attain, and hence we would expect the situation awareness of a passive observer to be fundamentally different in these circumstances. Evidence in support of a difference in situation awareness between active and passive involvement in a task comes from an unpublished study in the air traffic domain (Nuutilainen, 1997).

A further strategy is to collect responses to probe questions after a test trial. In this case it is more likely that information is retrieved from the long-term memory of what was previously in situation awareness, rather than the "current contents" in working memory. As might be expected, this approach yields a low correlation between situation awareness measures and prior performance (Sollenberger and Stein, 1995 in an air traffic control environment).

6.6.4.4 A summary of potential problems in using the SAGAT methodology

Practical issues

The approach requires a thorough analysis by HF and subject matter SME's to determine the range of information to be probed

A similar analysis is required to determine how to score probes for accuracy.

A large number of trials may be required to collect sufficient data to ensure reliable measures for each aspect of situation awareness probed.

Some pilot testing is required to determine the appropriate frequency and length of the freeze-probe stoppages. The goal is to ensure minimal primary task degradation, with a balanced trade-off between gathering multiple data points at each interval, versus the memory degradation that will occur if the probe interval is too long.

Validity Issues

There have been few full-scale studies that have undertaken a full validation of the methodology. In those studies where limited validation has been done, it is not clear whether the results will generalise to other application domains.

With extended exposure to the methodology, subjects may approach the task differently in order to better cope with having to satisfy the information demands from the probes. Hence, there may be a bias both in terms of the goals that the subject prioritises, as well as in the actual content of situation awareness compared with 'normal' behaviour.

There is a need to ensure that the probe questions samples broadly among information required for situation awareness.

In complex, dynamic, multi-tasking environments there is a high probability that freeze-probe interruption will cause a 'loss of the picture'. One approach to avoiding this problem is to use discrete trials of limited duration, with no continuity of the simulation in successive trials (e.g. Gugerty and Tirre, 1995a). However, this seems an impractical methodology to assess performance in a DMT context.

Implicit Measures

Implicit measures are sometimes referred to as "on-line indices"(Adams, Tenney and Pew, 1995) and imbedded or embedded measures. Whatever label is chosen, implicit measures cover a wide range of

measurement techniques, which all have in common the measurement of some aspect of the individual's *performance*. The actual outcome measured may be part of the normal behaviour associated with conducting a task, or it may result from an additional task, or probe, that the investigator introduces into the test environment. In many cases, investigators have tried to ensure that the latter are perceived as part of the repertoire of behaviours and responses that might be expected in a particular context. Thus, the widely adopted, divided attention task involving mental arithmetic while flying would ***not*** be an example of a suitable implicit performance measure. Whereas requesting a SitRep from a pilot by higher command would be. Embedded measures can be associated with the behaviours and outcomes illustrated in the "Performance" and "Behaviours" boxes of Figure 3.

The rationale underlying implicit measures is that they reflect directly some aspect of the individual's situation awareness. For example, if we were to monitor a pilot's waypoint precision while flying a mission, and then introduce into the scenario unexpected information that would affect performance (e.g. unexpected headwinds), we may infer that if the pilot adjusted thrust to compensate, then that information must have entered into the pilot's ongoing situation awareness at the appropriate point in time.

Of course, the problem for interpretation is what can be assumed if the pilot produces no increase in thrust. It would be dangerous to simply conclude the that the information never got into situation awareness, since the pilot may have ignored the information, or given it lower weight or decided that no response would be required at that moment in time. Hence, a generic problem for implicit measures to events introduced into a scenario, is how to interpret the absence of an expected response.

Since some degree of inference is required to interpret how intervening processes may have shaped the response outcome, implicit measures are seen to have greater reliability in ***experimental*** or ***simulation*** environments. This is because in these contexts the investigator can carefully control the timing, location and format of the information content to be probed. This control therefore allows some reasonable reduction in uncertainty about which ongoing tasks (and underlying information processes) are engaged at the time of the measurement.

To return to the general discussion of embedded measures, Adams, Tenney and Pew (1995) have suggested that on-line indices can be used to readily assess ease of processing, depth of processing and temporal processing of information. These concepts can be re-interpreted as correspondingly somewhat to Endsley' three levels of situation awareness. We will now examine, in turn, how implicit measures can be used for this purpose.

Implicit measures related to SA level 1 tell us how context influences detection accuracy and latency, particularly under conditions of divided attention. Probes can work in two ways in this context. First, by introducing information related to a new task that the individual must attend to, we can assess the manageability of the additional load by looking at how individuals respond to the added task demands. This approach could be used in a variety of contexts for examining the relative merits of different design options for the displayed information associated with the added task. In situation awareness terms it would be important to examine how the information format influences the ease of detecting new information (level 1) and ease of integration and comprehension into the current picture (level 2). The second way in which probes can be used to assess ease of processing of information, is to look at how long it takes for the individual to return to the primary tasks that were in place at the time of the interruption, having dealt with the additional task. This re-establishment of Global situation awareness is a major problem for pilots who have spent the past several minutes in an enemy engagement.

Implicit measures related to the *comprehension and integration* of information (level 2) can be implemented to address a number of fundamental issues. First, we can use the approach to examine how an individual is able to piece together separate information elements from different parts of the environment to reach a particular understanding. In particular, we can assess just how well some subtle pieces of information may be attended to and processed. That is, under some circumstances an individual must extract some very subtle cues from the environment that might influence the current understanding of a situation or prepare for some future action. Orasanu (1990) provides an excellent example in aviation context how to experimentally investigate the influence of subtle indicators in an aviation context.

Endsley (1995c) has criticised such measures somewhat severely. She notes that such manipulations of the external information environment may be obtrusive "*requiring the subject to undertake tasks involved with discovering what happened to the changed or missing data*". She adds that such manipulations may give rise to misleading results, if subjects then perform tasks in a different way from the norm or make assumptions about the environment that they would not typically do. This criticism however, is only valid if the experimental manipulation is *ecologically inappropriate*. By this we mean a transformation or change in the task environment that would not be normal or natural under any circumstances for that individual pursuing a specific task. Thus, if in a DMT, the experimenter introduced a digital display window in the periphery of the simulated windscreen and required the driver to press a button on the joystick (or do some other action) in response to a change in the displayed value, this would clearly change the pilot's normal approach to information processing and would be ecologically invalid. The point that Endsley seems to miss, is that in designing explicit measures the experimenter does not come up with just anything that will change the environment to see if it is noticed, but carefully selects the specific manipulation in terms of the individual's goal at hand and ongoing associated tasks. Thus, only goal and task-relevant manipulations should be considered for the purpose of measuring information acquisition and maintenance for situation awareness.

One consequence of taking this perspective, is that some existing measures of information processing, such as secondary tasks involving mental arithmetic and the like, cannot be used as surrogates for true measures of situation awareness. This is because such tasks involve non-ecological valid tasks for the pilot to perform. While they may provide some answers on the pilot's relative spare attentional or processing capacity in some contexts, they do not inform us in any way about the processes of building situation awareness and the ongoing state of comprehension.

As a footnote, it is somewhat ironic that Endsley criticises such explicit measures because they interrupt the subject's ongoing tasks and affect situation awareness. First, this is only true if the tasks are poorly chosen. Second, the same exact criticism can be applied to the situation awareness global assessment technique (SAGAT) which she has herself developed and which we describe in greater in the next section.

6.6.4.5 Literature supporting implicit measures

Support for performance based measures comes from work in an ATC environment (Pritchett, Hansman and Johnson, 1995), who recommended implicit measures for multitasking complex, dynamic environments, supplemented by knowledge-based measures of situation awareness as well as action outcomes. Brickman et al**(get ref follow up) have developed the Global Implicit Measure in the context of a redesign of aircraft avionics for improved situation awareness. They claim an advantage of this over SAGAT techniques because there is no interruption to task flow, and the method provides greater precision in measurement. In general, their technique combines the embedded measures

approach and broaden the scope of measurement by identifying a series of metrics that are appropriate for tapping into all relevant aspects of situation awareness for a specific context.

The validity and utility of embedded implicit measures has been analysed for an air traffic environment (Rantanen and Butler, 1995), who concluded that this approach was highly promising for diagnosing the state of the controller's situation awareness at any moment in time. The measures were seen as having greater diagnosticity than overall performance measures. The authors argue that such measures are to be preferred in dynamic contexts with high information rates and multitasking over post event probes, freeze techniques or secondary tasks.

6.6.4.6 A generic methodology for developing implicit measures

A general strategy for implementing this methodology by HF SMEs is summarised in point form below. Readers interested in obtaining more detailed information on the development of measures and methods for the assessment of situation awareness are referred to Matthews, Webb and McCann (1997) and Matthews, Webb and Bryant (1999).

1. Conduct an initial goal-oriented task or cognitive analysis to identify how the information required to build and maintain situation awareness varies over task contexts.
2. Assemble a scenario that contains the appropriate conditions to ensure that all of the goals and tasks of interest are systematically represented.
3. Identify the specific areas of situation awareness to be probed for each context, and the level of situation awareness that is to be the focus of interest.
4. Identify the technical feasibility for probe injection and performance measurement.
5. Establish valid and reliable performance measures for the behaviours to be sampled.
6. Determine a series of probes that can be embedded into the task structure to capture data for each metric.
7. Create a script that contains the temporal plan for injecting probes into the scenario and the capture of responses.
8. Create scoring criteria for the probe responses.
9. Determine the technical requirements for injecting real-time probes capturing on-line data. Establish alternate methodologies for response data that cannot be easily captured on-line (e.g. video tape).

6.6.4.7 Advantages and disadvantages of implicit measures

Advantages

There is no task interruption to contaminate situation awareness

Critical elements of the task context can be manipulated systematically

The subject's task has high ecological validity

Once created, scenarios can be re-run and modified relatively easily

The subject's responses to probes are part of the natural responses that would be normally made in the situation

All aspects of situation awareness can be readily assessed

Disadvantages

There may be some technical and logistical overhead in building a system that can render scenarios, introduce real time probes and capture on-line data.

Scenario generation and script building is time consuming.

Performance based measures, unless carefully chosen, may be too remotely linked to the underlying situation awareness factors that are supposed to be driving the performance outcome (Reising, 1995).

The absence of an expected response can be difficult to interpret.

Measures of accuracy and response time will require a significant number of trials (note that this is generally the case for all measures of situation awareness).

Subjective ratings of situation awareness.

In contrast with performance or probe-based measures, in this approach participants are required to generate numerical or scale ratings based upon their subjective evaluation of certain aspects of their situation awareness. Such ratings may be done during a brief interval while the mission scenario is temporarily frozen, directly after the event in question, or while viewing a scenario replay that captures the SA performance of interest.

An alternative approach that uses subjective assessment is to have the ratings done by SMEs of selected aspects of task performance.

Each of these methods will be reviewed below.

6.6.5 Subjective techniques: I. Post-event retrospective recall

In this approach individuals first perform a test session and then are probed for recall of certain information after the event. Clearly, while convenient to administer, a number of questions may be raised concerning the validity and reliability of this approach. First, if the test session is long subjects may have forgotten earlier information. Second, if the nature of the task environment is highly dynamic with changing information sources, quantity and type then a single post-event probe is unlikely to capture this complexity. Further, it is likely that under such circumstances the individual will have no explicit knowledge remaining of the information detail at any point in time. Given these limitations, it is possible that what a subject does when faced with this task is to make inferences about what her/his situation awareness might have been, rather than being able to draw on data directly that reveal what actually happened.

Interpreting the results of post-event measures is consequently problematic, since a low accuracy score could be as much due to a number of intervening variables as it would reflect actual situation awareness content. Therefore, except for short scenarios, in which the information content and sources to be monitored relatively low and only a single task is required, this procedure has few practical applications and would not be recommended for addressing situation awareness research questions in a driving context.

6.6.6 Subjective techniques: II. During task execution

In this procedure, the subject provides a numerical rating along a psychometrically valid scale that represents level of perceived situation awareness. The rating may be generated concurrently with the task at hand, if workload and information rates are low, however, under high rates of information and multitasking this technique may be too obtrusive. An alternate approach is to interrupt the ongoing

task and have the subject immediately generate a retrospective rating. If there are concerns with the impact of such task interruptions on the overall task performance and ability to maintain situation awareness, then the rating may be provided at the end of the task or at a convenient break in which task focus changes. The problem with such an approach is that the time period covered will be large and the validity of the rating provided may be low. If the situation awareness varied considerably over the period in question, it is not clear what a single numerical rating would represent or mean. To overcome this, it might be possible to videotape a subject during the course of a task, and then have them watch the playback and provide situation awareness ratings. These could be either qualitative in the form of a narrative of the content of their situation awareness from moment to moment, or quantitative using a rating on some specific aspect of situation awareness. The reliability, validity and diagnosticity of such an approach has yet to be demonstrated in the literature.

An example of a well-developed subjective rating method is the Situational Awareness Rating Technique (SART, Taylor, 1995) derived for use in the aviation domain. The method provides metrics for three basic dimensions of performance that the author believes contribute to situation awareness. These are demand (D) and supply (S) of attentional resources and (U) understanding of the situation. Each of these factors is in turn broken down into contributing dimensions as follows:

- *Attentional demand*: complexity, variability and instability
- *Attentional supply*: arousal, concentration, division of attention, spare mental capacity
- *Understanding*: information quality, information quantity, familiarity.

According to the model, scores on each of the sub-dimensions may be added to provide a score on each of the main dimensions. These scores may also be summed to provide an overall or global situation awareness metric.

Taylor's approach has been to provide psychometric validation for the measurement construct as opposed to validation against more direct measures of situation awareness. In this respect, the index was found to correlate with workload measures and to be sensitive to manipulations of task demand.

There appear to be several limitations with Taylor's approach.

The identified factors may or may not generalise directly to the DMT environment, nor describe it completely.

There is not theoretical or practical justification for simply adding the sub-component scores in a linear manner.

The level of diagnosticity provided by the method is yet to be validated, although claims are made that the overall situation awareness rating reflects gross aspects of task demand.

The simulation context in which the model was developed allowed did not provide a multi-tasking, high demand complex work domain and only gross aspects of the complex task of flying an aircraft were able to be represented.

The test/re-test reliability of SART has been found to be low (Vidulich, Crabtree and McCoy,1993).

Situation awareness is confounded with workload in the use of measures of attentional demand and supply.

6.6.6.1 SA-SWORD

This approached uses a modified version of the Subjective Workload Dominance (SWORD) technique and was developed to obtain subjective ratings of situation awareness of the information provided by

displays (Vidulich and Hughes, 1991). The method requires the subject to make a comparative evaluation of the relative amount of situation awareness provided by different display formats.

This technique would seem to have limited applicability in assessing SA in a DMT environment.

6.6.6.2 Advantages and disadvantages of subjective measures

Even when the circumstances are such that only a single subjective measure of situation awareness can be collected, there are some concerns about what exactly the ensuing number actually means. When a single metric is used it probably reflects some composite evaluation and integration of a variety of impressions that the subject has in any particular context. Hence, a single subjective measure may not be sufficiently diagnostic to provide a useful assessment of the particular content of situation awareness at any point in time. Also subjective ratings are unlikely to reflect any of the subjects gaps in situation awareness or failures to adequately integrate information. Subjective measures are often more likely to be measures of confidence in, rather than content of, situation awareness.

A number of problems in the use of global ratings have been reported in the literature. Small (1995) found that global ratings of situation awareness were found to be contaminated by other contextual, contributing factors. Small suggests that subjects may tend to rationalise prior behaviour when providing situation awareness ratings after the event, and ratings tend to be confounded by deficient or altered memory for the antecedent events. More problematic, Fracker (1991b) found that situation awareness ratings were uncorrelated with performance measures, a result that mirrors the finding for some global workload ratings (Yeh and Wickens, 1988).

In general, it would appear that only application environment where a single subjective rating may be useful is for comparisons of *overall* level of situation awareness in a gross manner from one context to another (like an overall mental workload measure). In this case, measures could not be compared across individuals or across situations in which goals and tasks differ.

In summary and conclusion of the usefulness and validity in using subjective ratings to assess situation awareness, we support Fracker's (1991b) following caution:

"...subjective measures may be measuring subjects' rational inferences about their situation awareness or workload rather than the results of their introspection. Until researchers have a better understanding of just how people produce their responses to subjective rating scales, caution in their use would seem to be in order".

6.6.7 Subjective techniques: III. SME ratings of behaviour

This method uses an assessment of pilot performance to assess SA by an SME who acts as an independent and trained observer. The methodology involves first identifying relevant behaviours that are thought to be highly dependent upon SA, second, selecting performance indicators of the behaviour, and third, developing psychometrically valid and reliable scales to score such behaviours. Conceptually, this approach shares much in common with objective performance measures to infer SA processes and many of the same tasks may be assessed using either approach, where feasible and appropriate. Embedded measures provide empirical data derived directly from performance. SME ratings provide data on an assessment of the performance. In either case, the measures are surrogates for SA, but have the advantage of being ecologically valid in that they tap into highly valid mission execution tasks that are dependent upon the quality of SA.

Clearly, subjective ratings of pilot behaviour may be open to contamination and lack the required levels of validity and reliability if they are not appropriately developed and applied. The literature

shows that asking open-ended questions such as "Did the pilot show the appropriate level of SA when engaging this threat" do not provide the psychometric precision and quality required. Too much is left for the rater to interpret, individual raters may differ in their interpretation of "appropriate", they may focus on different aspects of SA and the rater may not be able to be consistent across events or different individuals observed.

To overcome such problems with such rating techniques in assessing SA in air combat missions and training, the USAFRL has spent considerable time and effort in developing rating scales that have appropriate psychometric properties. These scales are generally referred to as Behaviourally Anchored Rating Scales (BARS) and use behavioural descriptors to ensure that scale points are used with precision and require minimum interpretation by the rater.

An example of a BARS for the mission execution phase and the identified behaviour of Navigation Accuracy is shown below. Note that this behaviour would have been selected by SMEs as being a critical task for mission execution that involves several aspects of SA.

MISSION EXECUTION				
1. Poor	**2. Marginal**	**3. Standard**	**4. Very Good**	**5.Exceptional**
NAVIGATION ACCURACY: AWARENESS OF CURRENT LOCATION, ADHERENCE TO PLAN CONSIDERED				
Crew is lost/disoriented No adherence to planning routing Unable to meet objective requirements	Often deviates from / is unsure of routing Only able to meet objective requirements with difficulty	Generally able to follow the planned routing Several large off-track deviations performed to meet objectives	Is able to adhere to planned mission routing Is aware of position with respect to objectives at all times	Is continually aware of position Able to make adjustments to meet objective requirements
Look for: Print (HC) cross-country ground track from IOS at each waypoint Note any crew discussions indicating confusion over present position Should stay within required (1.5mi) distance of each waypoint Perform adequate checks on nav data entered into flight computer				

Figure 4: Example mission execution BARS

Clearly, such a metric provides clear guidance for the rater in scoring the behaviours in question, even though phrases such as "often" and "generally" are employed.

USAFRL has encountered a number of problems in the conduct of such assessments and a number of recommendations from their literature (and others) are outlined below to enhance the reliability and validity of the data obtained.

Physiological measures

A comprehensive analysis and review of the applicability of physiological measures approaches to the assessment of situation awareness has yet to be conducted, particularly in terms of the reliability, validity and practicality. Possible measures identified include: EEG, heart rate, eye blinks, eye movement tracking, blood pressure, galvanic skin response and even newer technologies such as fMRI. In general, such measures will inevitably have limited applicability and practicality because of

the instrumentation requirements and restrictions in movement and posture that may be placed upon the normal behaviour of a pilot. Further, since such measures are global, and involve significant time lags between event antecedents and recorded responses, the face validity of their diagnosticity must be in question.

Wilson (1995) provides an illustrative study of how heart rate and eye blink measures might be implemented for assessing situation awareness in an aviation context. Unfortunately, Wilson concludes that a battery of such measures may be necessary, and hence their practicality for adoption in DMT environments is questionable.

The reader is directed to Vidulich, Stratton, Crabtree and Wilson (1994) for a more comprehensive review of a subset of physiological measures.

The most limiting factor concerning these measures is that, no matter how sophisticated, they fail to address the measurement of situation awareness with sufficient specificity to be useful. For example, knowing where a driver looks may provide information on whether appropriate scanning is taking place, but does not address whether information is detected and integrated with situation awareness. Similar arguments can be made against all of the physiological measures that have been tried or contemplated.

6.6.7.1 Summary evaluation of measures and recommendations

The following table represents our judgment of the various measures that have been used to assess situation awareness, in terms of validity, reliability, diagnosticity and practicality for use in the DMT context. For the most part, this judgment is based upon our own experience in using such measures and an analysis of the relevant literature. To date there is no definitive psychometric or other comprehensive evaluation that has been conducted to assess the relative merits of the different measurement approaches. The preferred measurement approaches that we recommend are shaded in grey.

Measure	Validity	Reliability	Diagnosticity	Practicality	Comments
Global outcome measures	High	Low	Low	High in simulator	Can be influenced by many extraneous variables
Explicit: task interruption and probe	Potentially high in moderately demanding tasks (as yet not empirically tested). Low in multi-tasking, highly dynamic environments	High-with sufficient trials	High-with sufficient probes/	Medium: requires large number of trials.	Requires careful analysis of goals and tasks May also transform task and take pilot out of the loop
Explicit: post event recall/probe	Low	Low	Low	High	Unsuitable for situations with highly dynamic and complex information content
Implicit measures	High	High-with sufficient trials	High	Medium: requires a controllable test environment	Feasible in a DMT context
Subjective ratings of situation awareness	Low	Medium	Low	High	Unsuitable for situations with highly dynamic and complex information content
Explicit: SME ratings using BARS	High	High with sufficient SMEs	Medium	High in simulator	A well tried technique and succssful technique in DMT
Physiological measures	Low	Unknown	Low	Low	

Table 9: Summary evaluation of situation awareness measures

Thus, we recommend two approaches to the measurement of SA during the execution phase of a mission scenario - implicit measures that generate objective performance data and SME ratings using BARS.

6.6.7.2 Detailed Recommendations for SA measures for different mission phases

In view of the specific mission tasking directed by the Scientific Authority to look at situation awareness, and its measurement, we have provide a more detailed outline of how this may be accomplished in the following paragraphs.

Based upon our review of the literature, understanding of mission contexts and discussions with SMEs, we are recommending that the following behaviours and performance indicators be the subject of assessment during mission execution as surrogate measures of SA.

In the tables that follow, EP refers to an Embedded Probe, PEMDA refers to Post-Event Mission Data Analysis and BARS to Behaviourally Anchored Rating Scales administered by SMEs. Both EP and PEMDA techniques rely mostly on Objective Performance Measures. The particular choice of measure will depend upon the logistical constraints and opportunities afforded by the particular mission simulation. Mission phases are abbreviated to: Start, taxi, take-off phase (STTO), transit to TRP (TTRP), Ingress/Execute/Egress (IEXEG)

We have categorised the measurement points according to the different types of SA as outlined previously, rather than organising them by the phases of a mission. Further, in recognition of the special local SA associated with air engagements, we have separated this out as a separate area for measurement.

Spatial, temporal and identity awareness: an appreciation of the spatial location of ownship, others in the formation, other relevant participating units, enemy formations and terrain. Knowledge of the changing spatial picture over time. Projecting flight vectors and trajectories. Temporal awareness is a key component for performing the mission in accordance with the plan and knowing when critical timing elements are jeopardised. the pilot's awareness of key navigational points and decision windows.

Function	Behaviour	Measure	Mission Phase	Method
Follow navigational plan	Planned waypoints followed	Spatial deviations from plan Temporal deviations from plan Communicates to crew deviations (spatial, time lateness) from plan	TTRP, IEXEG	PEMDA, BARS PEMDA, BARS PEMDA
	TRP on time	Adjust groundspeed/course to arrive at TRP on time	TTRP	PEMDA
	Crosschecks: Shows awareness of reduced/enhanced mission crosscheck time	Adjusts navigation/speed to compensate for deviation Communicates status to crew	IEXEG	EP, PEMDA
Maintain formation	Correct location in formation	Deviates significantly from formation to jeopardise mission success	TTRP, IEXEG	BARS
	Rejoins formation	Rejoins in a timely and positionally accurate manner	TTRP, IEXEG	BARS
Awareness of other Participating Units	Performs actions appropriate to awareness of other PUs	Communicates with PUs as required Communicates with crew re changes in PU status	TTRP, IEXEG	EP, BARS
Awareness of enemy positions	Updates plan in accordance with changes in enemy positions	Changes altitude/course/speed appropriately to reflect changes in enemy positions Communicates to crew info concerning changes	TTRP, IEXEG	EP, PEMDA, BARS

Table 10: Spatial, temporal and identity awareness

System awareness: attention to the relevant information from aircraft systems (e.g. flight dynamics, weapons, navigation, sensor).

Function	Behaviour	Measure	Mission Phase	Method
Systems status (weapons, sensors, fuel, navigation and power systems etc)	Performs system check	Completes all items of check. Notices problem areas Checklist completed at required time	ALL	EP, PEMDA, BARS
	Directs formation re checks	Pilot calls for crew to perform appropriate checks at required time	ALL	PEMDA, BARS

Table 11: System awareness

Resource and crew awareness: in a DMT environment this would involve the monitoring of other units of the formation and comprehending how changes in their status and capability impact upon ownship/formation ability to meet mission goals.

Function	Behaviour	Measure	Mission Phase	Method
Changes in crew status	Recognises changes in crew status that impact upon mission execution	Visually identifies changes in crew status Responds to communications concerning crew status	ALL	EP, PEMDA, BARS
	Adjusts plan in accordance with changes in crew status	Revises navigational plan Communicates change in plan	ALL	PEMDA, BARS
Formation integrity	Crew out of formation	Detects out of formation status and communicates to crew	ALL	PEMDA, BARS

Table 12: Resource and crew awareness

Tactical awareness during engagement: detects contacts to be engaged, positions aircraft in accordance with tactical plan, recognises enemy tactics, co-ordinates with formation.

Function	Behaviour	Measure	Mission Phase	Method
Contact initial engagement	Detects and recognises contacts	Manoeuvres a/c in accordance with plan Communicates to formation	IEXEG	EP, PEMDA, BARS
Spatial awareness during the engagement	Recognises enemy tactics	Positions a/c in accordance with the plan and enemy actions	IEXEG	EP, PEMDA, BARS
Co-ordinates with formation	Communicates intent, expectations	Timely and accurate communications to formation	IEXEG	PEMDA
Monitors formation	Formation elements acting according to plan	Recognises anomalies in formation engagement plan Recognises formation out of position Communicate to formation anomalies in engagement plan	IEXEG	EP, PEMDA, BARS

Table 13: Tactical awareness during engagement

Mission/goal awareness: At the highest level it will be the tactical mission objectives for that unit (e.g. ingress, deliver ordinance, egress, deal with threats); updates to mission goals as a result of changes in situation; rejoins mission plan following an engagement

Function	Behaviour	Measure	Mission Phase	Method
Re-engages primary mission goals after an engagement	Rejoins formation	Joins formation in accordance with plan	IEXEG	PEMDA, BARS
	Adjust navigational plan	Changes course/speed to bring position back into line with plan	IEXEG	PEMDA, BARS
Assesses and responds to changes in mission picture	Acquires and processes information to update status of mission	Communicates with formation and other units to acquire and disseminate knowledge	IEXEG	PEMDA, BARS
	Updates plan to confirm to changed mission picture	Changes course/speed to reflect new plan Communicates changed picture and plan	IEXEG	PEMDA, BARS

Table 14: Mission/goal awareness

7. The complete measurement model

In this section we bring together the behaviour domains and behaviour constructs with the MOPs outlined above in order to provide a complete and integrated overview of the recommended measurement approach.

For convenience, the measurement domain is broken down into the components of mission briefing, execution and debriefing in the following tables. Note that OP=Objective performance measures, BARS=Behaviourally Anchored Ratings Scales and PF=Pathfinder. The numbers in the columns are cross referenced to Annex A, where the detailed MOPs are listed using the structure of the following tables and are separated into two sections- BARS and Objective Performance measures.

Behaviour Domains	Behaviour Categories	Specific Behaviours	Measures		
			OP	BARS	PF
Detailed Planning	Tactics	Task understanding	1A	1A	
		ROE understanding	1A	1A	
		Route review/analysis		1A	
		Tactical effectiveness of plan		1B	
		Factors considered in plan		1C	
		COA considered	1B, C	1C	
		Decision quality/timeliness		1D	
		Use of resources		1E	
		Creation of extra materials		1F	
	Time Management	Time appreciation	2A—	2	
		Time required	2E	2	
		Efficiency		2	
	CRM-Function Allocation	Clarity and assignment of team roles	3A, 3B	3	
Planning Products		Quality of:			
		Fuel plan		4A	
		TOLD		4B	
		Communications		4C	
		Mission data		4D	
Briefings	Communication	Detail	4A—	5	
		Participation	4D	5	
		Comprehensiveness		5	
		Time required		5	
	General effectiveness	Degree of instructor intervention		6	
Mental Model of mission	Knowledge Structures	Change in KS resulting from training			•

Table 15: Measurement Concept: Planning and Briefing

Behaviour Domains	Behaviour Categories	Specific Behaviours	Measures		
			OP	BARS	PF
Mission conduct	Plan compliance	Navigation accuracy Time control Communications required Compensation/adjustment Loss of separation incidents	5A 5B,C 5D 5E 5F	7A 7B	
	Communications	Speed/accuracy Terminology Necessity Information content	6A-E 6F,G 6H 6I	8	
	Aircraft handling/control	Piloting skills (e.g accuracy in heading, speed, altitude, clearances)		9	
Situation awareness	System awareness	Checklist performance	7A-7D	10A	
		Sensors, status indicators	8A,B	10B	
	CRM	Crew awareness and communication	9A-E	11	
	Tactical awareness	Contact detection(speed,accy) Spatial awareness Co-ordination COA	10A-D 10E-I	12 12 12 12	
	Mission/goal awareness	Resumption of plan after engagement Detects changes in mission picture Changes/updates plan re changes in mission picture Communicates with crew	11A 11B 11C 11D	13 13 13 13	
	Flying skills	Engagement skills Weaponry skills Threat exposure Role discipline		14	
Goal Accomplishment	Achievement of objectives	Completeness, survival		15	
	General effectiveness	Degree of instructor intervention		16	
Mental Model	Knowledge structures	Team mental model/co-ordination/CRM			■

Table16: Measurement Concept: Mission Execution

Behaviour Domains	Behaviour Categories	Specific Behaviours	Measures		
			OP	BARS	PF
Briefings	Communication	Detail, Participation, Comprehensiveness, Time spent	12A	17	
	Mission accomplishment	Time spent Completeness	12B	18	
	Crew CRM	Time spent Completeness	12C	19	
	Crew Technical	Time spent Completeness	12D	20	
	Lessons Learned	Time spent Relevancy	12E	21	
	General effectiveness	Degree of instructor intervention	12F	22	

Table 17: Measurement Concept: Debriefing

THIS PAGE INTENTIONALLY LEFT BLANK

8. Measurement procedures and logistics

We assume that the general measurement context will be one that allocates a reasonable amount of time to mission planning, execution and debriefing functions. In the case of the application of the Pathfinder methodology to assess changes in knowledge structures, it is critical that the test environment provides a context where such changes are likely to occur. This could be in the mission planning phase if sufficient time is given to allow teams to develop a clear and detailed mental model of the operational plan and procedures. However, in order to ensure that there was every opportunity for a change in knowledge structures to occur several precautionary steps would need to be taken. First, participants should have equivalent experience and knowledge states at the start of the mission planning. Hence, crew who had done this before or were familiar with the mission scenario and plans from other contexts would not be suitable. Second, objective training goals or end states should be identified to ensure that concepts to be acquired during training can be assessed using the technique. Third, time must be found and allocated to ensure that the participants have sufficient time (about 45 minutes to an hour) to complete the rating procedure at the start and end of mission planning. Thus, because of these constraints, it may be more realistic to consider the first pilot study involving Pathfinder to be conducted in a training environment that affords more flexibility than the typical DMT trial. For example, using pilots who have come out of initial training and are starting the CF-18 course.

8.1 Data Capture

For the application of the objective performance measures and BARS scales there are a number of similar requirements that are necessary for data collection in all mission phases. In general, because of the complexity of the behaviours, communications and events it may not be possible to assess all aspects of performance by live observers in real time. Therefore the preferred method is to create an accurate audio and video data record of all of the behaviours and events of interest and then to analyse this after the trial. This approach ensures an initial higher accuracy of data capture, especially under highly dynamic conditions, and more time for detailed analysis and assessment of events that may be complex and involve dynamic interactions between participants. The exception to this might be the planning and debriefing stages of a mission, whereby live observers (trained SMEs) can adequately keep up with the data rates and simultaneously capture the measures of interest and score the behaviours. USAFRL experience suggests that this can be done in real time, but we have not been able to assess the extent to which behaviours may have been missed or inaccurately scored.

Further requirements for the planning phase are that all planning products be available for analysis by assessors and that all communications between distal participants can be recorded and logged (time, sender, receiver, content) on a common time base with the local communications that are taking place. The availability of the teleconference record should facilitate this process.

For the mission phase, we have assumed that real-time data capture of all critical events, communications, crew behaviours, instrumentation readings and external environment are captured by the simulation software and are subsequently available for analysis. In particular, for DMT it will be important to have a common timebase and record of all relevant communications from disparate and distal players and participants.

8.2 Personnel

Some effort and resources will need to be expended in selecting and training the SMEs who will be required for administering BARS scales and evaluating objective performance data. Clearly, the most obvious personnel will be AF trainers and DMT staff responsible for setting training goals, designing missions and evaluating personnel performance.

8.3 When to Measure?

Spiker, Tourville, Silverman and Nullmeyer (1996) developed a conceptual model of a mission that can be applied to most weapons systems. This detailed model serves the detailed measures applied at USAFRL. This model serves as the basis for a more generic model to be used in Canadian DMT. This model borrows from work done by TopACES (2003) and is presented in Figure 1.

This conceptual model of a mission is described at a level that is generic and applicable to most missions and weapons systems. This model can then be used to illustrate when, in the course of a DMT simulation, measurement activities can occur. Whether or not measurement can take place depends upon the degree to which the participants need to be aware of the measurement and the training activity in which they are engaged. If the measurement requires questionnaires or some other sort of participant response, measurement will need to take place between phases (in order that the participant's cognitive process is not interrupted). If the measurement relies on SME observation it can be undertaken at any point, assuming that the simulator is such that the SME can adequately and unobtrusively observe the participant. If the measurement is an event-based metric that has corresponding automated data capture, measurement can occur at any point.

Within the context of this project, behaviourally-based measures of mission planning, mission briefing, mission rehearsal and mission debriefing can occur during those phases of the mission. This assumes that a SME is assessing performance in line with a behaviourally-based listing of criteria and the participants themselves act as they normally would during these activities.

Event-based measures of Situation Awareness can occur during the more tactical phases of the mission, notably STTO, TRP, ingress, attack and egress. Event-based measures are predefined and the specific triggers may be inserted into the training scenario. Automated data collection can be used to collect the desired response, which must be subsequently evaluated by an SME. Finally, Pathfinder measurement of mental model development occurs before the mission begins and after it finishes. This cannot occur during the mission as it requires the participants to complete a concentrated rating phase.

8.4 Who to Measure?

Who to measure will depend upon the extent to which participants are involved. For instance, personnel who are familiar with the simulator or the scenario (e.g. games controllers and other similar support staff) would likely skew the results if they were measured. However, their understanding of the plans of the 'real' participants, as gained through briefing, may be a good indicator of briefing performance. These personnel can perhaps be used as SME raters.

Mission training involves teams and attempts to provide individual practice and foster a mutual understanding of the goals of others and the means by which they will achieve those goals. By appreciating the mental model used by others, it is far more likely that a team member will be able to anticipate their needs and actions and, in so doing, support them effectively.

Measurement should therefore be made of all those members of the team deployed specifically to participate in the simulation exercise. Individuals whose main role it is to maintain and operate the simulator should not be measured, although they can be used as SMEs for specific measurement activities. Table 17 shows support staff and training participants against the phases in the conceptual model of a mission. Measurement should be focused on the training participants, with the possibility of using support staff as measurement SMEs (based on their experience with the simulator, the scenarios, and the possible outcomes).

Mission Phase	Support Staff Involement?[3]	Training Participant Involvement?[4]	Measurement Point?
Mission Planning	Yes	CF-18/AWACS Controller/Other Friendly Force Pilots	Pathfinder measurement of mental models BEFORE phase Behaviour-based measurement by SME during phase
Mission Rehearsal	Yes	CF-18	Behaviour -based measurement by SME during phase
Mission Briefing	Yes	CF-18/Other Friendly Force Pilots	Behaviour -based measurement by SME during phase
STTO	Yes	CF-18	Event-based measurement of SA during phase
TRP	yes	CF-18/AWACS Controller	Event-based measurement of SA during phase
Ingress	Yes	CF-18/AWACS Controller/Other Friendly Force Pilots	Event-based measurement of SA during phase
Attack	Yes	CF-18/AWACS Controller/Other Friendly Force Pilots	Event-based measurement of SA during phase
Egress	Yes	CF-18/AWACS Controller	Event-based measurement of SA during phase
Mission Debriefing	yes	CF-18/AWACS Controller/Other Friendly Force Pilots	Pathfinder measurement of mental models AFTER phase

Table 18: Measurement Points and Simulation Participants

From the table above, it is apparent that all simulation participants should be measured at all the identified measurement points (in the context of this MOP scheme) to ensure that all participants are realising some benefit from the simulation.

[3] Support Staff (devising and preparing scenarios, driving simulator entities, etc.) operate at all stages of the mission, and before participants arrive, and after they leave.

[4] Column is from the perspective of the CF-18 pilot; therefore CF-18 pilot always appears, but the appearance of AWACS or other pilots depends upon whether they will be involved with the CF-18 pilot during that mission phase.

THIS PAGE INTENTIONALLY LEFT BLANK

9. Measurement Conditions

Opportunities for conducting assessment are generally expected to be associated with regularly scheduled AF training courses, readiness training exercises, mission feasibility assessments and DMT exercises.

It is assumed that the major decisions concerning scenario details will be made by DMT or other training personnel responsible for setting training goals. Ideally, to maximise the chances of ensuring performance variance that can be captured by the MOPs the scenario should have elements of time pressure, real-world limitations and relevance, involve a high degree of co-ordination among crew and supporting units, have specific tactics to be adopted and uncertainty concerning threats and/or intelligence.

9.1 Data collection

The following issues need to be considered for data collection, whether done in real-time or post-event:

- Specific dates and times for data capture
- Liaison and points of contact between contractors and AF personnel
- The particular training scenarios to be employed
- The trial participants (experience, number, other limiting selection factors)
- Logistical support otherwise not provided in the trial environment (e.g. provision of technicians to address equipment problems, trial co-ordinator, data collectors)
- Methods for data handling and backup
- Training of data recorders and analysts
- Briefing of all trial administrators and relevant personnel

9.2 Data Analysis

It is assumed that the data analysis will be conducted by the contractor responsible for the trial administration with assistance provided by SMEs where required. The locus and means of the data analysis will be highly dependent upon the trial circumstances. Normally BARS and other data captured by video or audio recording may be taken off site for local analysis in a secure facility. However, where the data can only be replayed with specific equipment that is only available at the trial site, then arrangements will need to be made to ensure that data analysts have full and unhindered access to the data records.

To maximise efficiency of post-event analysis of data records, the technology should provide analysts with the following capabilities:

- Co-ordinated replay of all relevant data (e.g. communications, external environment, instrumentation, crew actions) on a common time base

- Ability to slow, pause, speed-up, backup and fast forward the data record
- An ability to identify source of data (e.g. who sent comm, who received)

A preliminary data reduction will be performed to extract from the raw data the specific MOPs for behaviour of interest. These data can then be coded into a spreadsheet or database and then taken to a secure site for further analysis, thereby freeing up the equipment and resources at the training site associated with data replay.

10. Conclusions

The goals of this project as outlined in the Statement of Work, and elaborated upon in subsequent discussions with the Scientific Authority, were to review the literature and develop MOP options for evaluating DMT and to develop a plan for a future Pathfinder trial to assess changes in structural knowledge resulting from AF training.

After reviewing and analysing over sixty papers, and having detailed discussions with SMEs, we have been able to deliver almost 100 practical MOPs (excluding Pathfinder analysis) broken down as follows:

	BARS	Objective Performance
Mission Planning	20	15
Mission Execution	12	39
Mission Debriefing	6	6

Obviously, it is not the recommendation that all of these MOPs be used, instead the inventory provides a source from which specific measures van be selected according to the measurement goals and local logistics of conducting the measurement.

The BARS MOPs are largely based on material that has been tried and validated at USAFRL and have been extended and modified to apply to a Canadian context and to mission aspects, such as situation awareness. In extending and adapting the measures and developing Objective Performance measures, we have relied extensively upon our own experience in measuring operational performance in other military contexts supplemented by AF SME input from Top Aces Consulting.

In addition to the MOP recommendations, we have also delivered a detailed trial plan for the application of Pathfinder, as requested, and this can be found in Annex B. The costing for the plan has been delivered separately to the Scientific Authority in electronic format.

THIS PAGE INTENTIONALLY LEFT BLANK

11. References and sources consulted

(2000). CF-18 Learn to Fly Manual, Canadian Air Force.

(2001). Trial VirtEgo - Timetable for White Force. Aircrew and Trials Teams: 14.

(2002). ACM 1 Briefing Presentation.

ADAMS, M. J., Y. J. TENNEY and R. W. PEW (1995). "Situation awareness and the cognitive management of complex systems." Human Factors 37(1): 85-104.

BENNETT JR., W., P. CRANE, E. SMITH, H. MCINTYRE, S. GRANT, I. MACK and F. LICHACZ (2002). TTCP HUM TP2 & AER TP1 Collaborative activity coalition mission training research: Situation awareness topics for discussion at USFRL. Meza, AZ: 4.

BENNETT JR., W., B. T. SCHREIBER and D. H. ANDREWS (2002). "Developing competency-based methods for near real-time air combat problem solving assessment." Computers in Human Behaviour 18: 773-782.

BERGONDY, M., J. E. FOWLKES, J. GUALTIERI and E. SALAS Key team competencies for navy air wings: a case study.

BORING, R. (2001). Adding Pathfinder to your HCI tool chest. Ottawa, ON, Carlton University.

BRICKMAN, B. J., L. J. HETTINGER, M. M. ROE, D. STAUTBERG, M. A. VIDULICH, M. W. HAAS and R. L. SHAW (1995). An assessment of situation awareness in an air combat simulation: The global implicit measure approach. International Conference on Experimental Analysis and Measurement of Situational Awareness, Daytona Beach, FL, Embry-Riddle Aeronautical Press.

CASTILLO, A., W. BENNETT JR., B. WENZEL, M. PARK, R. SCHVANEVELDT, R. ROBBINS, J. WOOSTER and S. KOTTE (2002). An innovative approach for assessing knowledge in air-to-air distributed mission training: 9.

COOKE, N. J. (1990). Using Pathfinder as a knowledge elicitation tool: link interpretation. Pathfinder Associative Networks: Studies in Knowledge Organization. R. SCHVANEVELDT. Norwood, NJ, Ablex.

DEARHOLT, D. W. and R. SCHVANEVELDT (1990). Properties of Pathfinder Networks. Pathfinder Associative Networks: Studies in Knowledge Organization. R. SCHVANEVELDT. Norwood, NJ, Ablex.

DENNING, T., W. BENNETT JR. and P. CRANE (2002). Mission complexity scoring in distributed mission training. Interservice/Industry Training Systems and Education Conference.

DUNLAP, J. C. and S. GRABINGER (1994). Using pathfinder networks to examine structural knowledge. Annual Meeting, 3rd International Interdisciplinary Conference of the International Research Network of Training and Development Conference, Milan, Italy.

DURSO, F. T. and K. A. COGGINS (1990). Graphs in the social and psychological sciences: empirical contributions of Pathfinder. Pathfinder Associative Networks: Studies in Knowledge Organization. R. SCHVANEVELDT. Norwood, NJ, Ablex.

DWYER, D. J., R. L. OSER, E. SALAS and J. E. FOWLKES (1999). "Performance measurement in distributed environments: intial results and implications for training." Military Psychology 11(2): 189-215.

ENDSLEY, M. R. (1988). Design and evaluation for situation awareness enhancement. Riding the Wave of Innovation, Human Factors Society 32nd Annual Meeting, Anaheim, CA, Human Factors Society.

ENDSLEY, M. R. (1995). Theoretical underpinnings of situation awareness: a critical review. Experimental Analysis and Measurement of Situational Awareness. D. J. Garland and M. R. Endsley. Daytona Beach, FL, Embry-Riddle Aeronautical Press: 17-23.

ENDSLEY, M. R. (1995). "Toward a theory of situation awarness in dynamic systems." Human Factors 37(1): 32-64.

ENDSLEY, M. R. (1995). "Measurement of situation awareness in dynamic systems." Human Factors 37(1): 65-85.

ENDSLEY, M. R. (1996). Situation awareness measurement in test and evaluation. Handbook of Human Factors Testing and Evaluation. G. O. O'Brien and S. G. Charlton. Mahwah, NJ, Lawrence Erlbaum Associates.

FRACKER, M. L. (1990). Attention gradients in situation awareness. Situational Awareness in Aerospace Operations. Neuilly-sur-Seine, France, AGARD. AGARD-CP-478: 6/1-6/10.

FRACKER, M. L. (1991). Measures of situation awareness: an experimental evaluation. Wright-Patterson AFB, OH, Armstrong Laboratory.

GENTNER, F. C., P. H. CUNNINGHAM and W. BENNETT JR. (1998). Integrated taxonomy to assess wayfighting effectiveness and human performance readiness. International MTA, Pensacola, FL, IMTA.

GENTNER, F. C., T. C. TILLER, P. H. CUNNINGHAM and W. BENNETT JR. (1999). Scenario- and behaviorally-based metrics for evaluating warfighter performance in distributed mission training. 41st Annual Conference of the International Military Testing Association (IMTA), Monterey, CA, IMTA.

GENTNER, F. C., T. C. TILLER and P. H. CUNNINGHAM (1999). Using mission essential MOEs/MOPs for evaluating effectiveness of distributed mission training. Interservice/Industry Training, Simulation and Education Conference, I/ITSEC.

GOLDSMITH, T. E. and P. J. JOHNSON (1990). A structural assessment of classroom learning. Pathfinder Associative Networks: Studies in Knowledge Organization. R. SCHVANEVELDT. Norwood, NJ, Ablex.

GOLDSMITH, T. E. and D. M. DAVENPORT (1990). Assessing structural similarity of graphs. Pathfinder Associative Networks: Studies in Knowledge Organization. R. SCHVANEVELDT. Norwood, NJ, Ablex.

GUGERTY, L. and W. TIRRE (1995). Comparing explicit and implicit measures of situation awarness. Experimental Analysis and Measurement of Situational Awareness. D. J. Garland and M. R. Endsley. Daytona Beach, FL, Embry-Riddle Aeronautical Press: 259-265.

GUGERTY, L. and W. TIRRE (1999). Individual differences in situation awareness. Situation Awareness Analysis and Measurement. D. J. Garland and M. R. Endsley. Mahwah, NJ, Erlbaum.

HOLLOWELL, R. and D. SVAB (2002). C-130 CRM Process Worksheet, Version 2.0, Air Force Research Laboratory: 12pp.

JENSEN, R. P. (1999). Human-Centered Development for Distributed Mission Training Systems. Annual I/ITSEC Conference, Orlando, FL.

KARP, M. R., D. CONDIT and R. T. NULLMEYER (2000). Cockpit/crew resource management for single-seat fighter pilots: 10.

KOLITZ, S. E. and R. M. BEATON (1993). Overall system concepts in mission planning. AGARD Lecture Series on New Advances in Mission Planning and Rehearsal Systems, AGARD.

LANDY, F. J. and J. L. FARR (1980). "Performance Rating." Psychological Bulletin 87: 72-107.

MARSHAK, W. P., G. KUPERMAN, E. G. RAMSEY and D. WILSON (1987). Situational awareness in map displays. Rising to New Heights with Technology, Human Factors Society 31st Annual Meeting, New York, NY. Human Factors Society.

MATTHEWS, M. L., R. D. G. WEBB and C. MCCANN (1997). A Framework for Evaluation of MIlitary Command and Control Systems. Toronto, DRDC.

MATTHEWS, M. L., R. D. G. WEBB and D. J. BRYANT (1999). Cognitive Task Analysis of the HALIFAX Claass Operations Room Officer. Toronto, DRDC.

MATTHEWS, M. L., L. E. BRUYN, A. R. KEEBLE and R. D. G. WEBB (2003). Assessing the Impacat of Multi-Source Data Fusion on Command and Control Operations in the HALIFAX Class Frigate: Use of the Operations Room Team Trainer (ORTT) Training Records to Extract Quantitative and Diagnostic Measures of Performance. Toronto, DRDC.

MEISTER, D. (1985). Behavioural Analysis and Measurement Methods. New York, NY, Wiley.

MORRISON, J. E. and L. L. MELIZA (1999). Foundations of the after action review process.

NULLMEYER, R. T. CRM Skills and Air Force Fighter Pilots, Air Force Research Laboratory: 16pp.

NULLMEYER, R. T., P. CRANE, G. D. CICERO and V. A. SPIKER (2000). A bridge between cockpit/crew resource management and distributed mission training for fighter pilots: 12.

NULLMEYER, R. T. and A. SPIKER (2000). Simulation-based mission rehearsal and human performance. Aircrew training and assessment: methods, technologies, and assessment. D. H. A. H.F O'Neil Jr., Lawrence Erlbaum Associates: 15pp.

NULLMEYER, R. T. and V. A. SPIKER (2002). Exploiting archival data to identify CRM training needs for C-130 aircrews: 11.

NULLMEYER, R. T. (2003). Main determinants of mission success. T. M. Lamoureux.

NUUTILAINEN, P.R.K. (1997). The effects of automation on situation awareness in an air traffic control simulation. Thesis, Department of Psychology, University of Guelph: 62pp.

OLMOS, O., C. LIANG and C. D. WICKENS (1995). Construct validity of situation awareness measurements related to display design. Experimental Analysis and Measurement of Situational Awareness. D. J. Garland and M. R. Endsley. Daytona Beach, FL, Embry-Riddle Aeronautical Press: 219-225.

ORASANU, J. M. (1990). Shared mental models and crew decision making. Princeton, NJ, Princeton University, Cognitive Sciences Laboratory.

PEW, R. W. (1994). An introduction to the concept of situation awareness. Situational Awareness in Complex Systems. R. D. Gilson, D. J. Garland and J. M. Koonce. Daytona Beach, FL, Embry-Riddle Aeronautical Press: 17-26.

SALAS, E., C. PRINCE, C. A. BOWERS, R. J. STOUT, R. L. OSER and J. A. CANNON-BOWERS (1999). "A methodology for enhancing crew resource management training." Human Factors 41(1): 161-172.

SARTER, N. B. and D. D. WOODS (1991). "Situation awareness: a critical but ill-defined phenomenon." International Journal of Aviation Psychology 1(1): 45-57.

SCHREIBER, B. T., R. WATZ and W. BENNETT JR. (2002). Development of a distributed mission training automated performance tracking system: 11.

SCHVANEVELDT, R. and F. T. DURSO (1981). General Semantic Networks. Annual Meeting of the Psychonomic Society, Philadelphia, PA.

SCHVANEVELDT, R., F. T. DURSO, T. E. GOLDSMITH, T. J. BREEN, N. M. COOKE, R. TUCKER and J. C. DEMAIO (1985). "Measuring the structure of expertise." International Journal of Man-Machine Studies(23): 699-728.

SCHVANEVELDT, R. (1990). Pathfinder associative networks: studies in knowledge organisation. Norwood, NJ, Ablex.

SCHVANEVELDT, R., R. TUCKER, A. CASTILLO and W. BENNETT JR. (2001). Knowledge acquisition in distributed mission training. Annual I/ITSEC Conference, Orlando, FL.

SHEEHY, E. J., E. C. DAVEY, T. T. FIEGEL and K. Q. GUEO (1993). Usability benchmark for CANDU annunciation - lessons learned. ANS Topical Meeting on Nuclear Plant Instrumentation, Control and Man-Machine Interface Technology, Oak Ridge, TN.

SMALL, S. D. (1995). Measurement and analysis of situation awareness in anesthesiology. International Conference on Experimental Analysis and Measurement of Situational Awareness, Daytona Beach, FL, Embry-Riddle Aeronautical Press.

SMITH, K. and P. A. HANCOCK (1994). Situation awareness is adaptive, externally directed consciousness. Situational Awareness in Complex Systems. R. D. Gilson, D. J. Garland and J. M. Koonce. Daytona Beach, FL, Embry-Riddle Aeronautical Press: 59-68.

SMITH-JENTSCH, K. A., J. H. JOHNSTON and S. C. PAYNE Measuring team-related expertise in complex environments: 61-87.

SMITH-JENTSCH, K. A., R. L. ZEISIG, B. ACTON and J. A. MCPHERSON Team dimensional training: a strategy for guided team self-correction: 271-297.

SMITH-JENTSCH, K. A., D. M. MILANOVICH and D. C. MERKET (2001). Guided team self-correction: a field validation study. Enhancing team performance: emerging theory, instructional strategies, and evidence. 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

SPIKER, V. A. and R. T. NULLMEYER (1995). Benefits and limitations of simulation-based mission planning and rehearsal. Eighth International Symposium on Aviation Psychology, Ohio State University.

SPIKER, A. and R. T. NULLMEYER (1995). Measuring the effectiveness of mission preparation in the special operations forces. Brooks Air Force Base, Texas, Aircrew Training Research Division: 118pp.

SPIKER, A., S. J. TOURVILLE, D. R. SILVERMAN and R. T. NULLMEYER (1996). Team performance during combat mission training: a conceptual model and measurement framework. Mesa, AZ, United States Air Force Armstrong Laboratory: 64pp.

SPIKER, A., D. R. SILVERMAN, S. J. TOURVILLE and R. T. NULLMEYER (1997). Tactical team resource management (T2RM) effects on combat mission training performance. Mesa, AZ, Aircrew Training Research Division, Airforce Research Laboratory.

SPIKER, V. A., S. J. TOURVILLE, M. J. BRAGGER, T. S. G. D. DOWDY and R. T. NULLMEYER (1999). Measuring C-5 crew coordination proficiency in an operational wing. 21st Interservice/Industry Training Systems and Education Conference, Orlando, FL.

SPIKER, V. A., R. T. NULLMEYER and S. J. TOURVILLE (2000). Relationship between mission preparation and performance during combat mission training: 10.

SPIKER, A., R. T. NULLMEYER and S. J. TOURVILLE (2001). Impacts of mission preparation on mission performance. Interservice/Industry Training, Simulation and Education Conference, I/ITSEC.

STOUT, R. J., J. A. CANNON-BOWERS, E. SALAS and D. M. MILANOVICH (1999). "Planning, shared mental models, and coordinated performance: an empirical link is established." Human Factors 41(1): 61-71.

TANNENBAUM, S. I., K. A. SMITH-JENTSCH and S. J. BEHSON Training team leaders to facilitate team learning and performance: 247-270.

TAYLOR, R. M. (1993). Human factors of mission planning systems: theory and concepts. New Advances in Mission Planning and Rehearsal Systems. AGARD.

THOMPSON, J. S., S. J. TOURVILLE, V. A. SPIKER and R. T. NULLMEYER (1999). Crew resource management and mission performance during MH-53J combat mission training. 21st Interservice/Industry Training Systems and Education Conference, Orlando, FL.

THURMAN, R. A. and R. D. DUNLAP (1999). Assessing the effectiveness of simulator-based training. Interservice/Industry Training, Simulation and Education Conference, I/ITSEC.

TOP ACES INC. (2003). Situational Awareness Report. Toronto, ON, Defence Research and Development Canada.

TOP ACES INC. (2003). Experimental Protocol Design - SA. Toronto, ON, Defence Research and Development Canada: 9pp.

USAF (1998). Air Force Task List (AFTL): Air Force Doctrine document 1-1, United States Air Force.

USAF (2002). C130-CRM Process Worksheet. Mesa, AZ, Air Force Research Laboratory.

USAF (2002). C130 Mission Performance Worksheet. Mesa, AZ, Air Force Research Laboratory.

VIDULICH, M. A. and E. R. HUGHES (1991). Testing a subjective metric of situation awareness. Visions, Human Factors Society 35th Annual Meeting, San Francisco, CA, Human Factors Society.

VIDULICH, M. A., M. CRABTREE and A. L. MCCOY (1993). Developing subjective and objective metrics of pilot situation awareness. 7th International Symposium on Aviation Psychology, Columbus, Ohio State University.

VILLACHICA, S. (1999). An Investigation of the Reliability of Pathfinder Networks, University of Northern Colorado.

WHITAKER, L. A. and G. A. KLEIN (1988). Situation awareness in the virtual world: situation assessment report. 11th Symposium on Psychology in the Department of Defense.

Annex A: Recommended Measures of Performance (MOPs)

This Annex is intended as a standalone guide for determining appropriate MOPs for various aspects of a mission. The information is organised in terms of the major functional sequences of the mission, namely planning, execution and debriefing. In general the MOPs are based upon recommendations from the literature reviewed or based upon HSI's experience in C2 measurement in other domains. A multi-measure, multi-method approach is recommended to maximise validity and reliability of the data obtained. For a more detailed explanation of the derivation and context for the measures, readers should consult the main body of the report.

It should be noted that the examples provided below are illustrative and representative of the measures to be employed. Thus, additional detail and further items for any particular scale or measure will need to be refined and applied to a Canadian context. For example, we may illustrate a behaviourally based assessment of "time management" during the planning process with one or two sample rating items. The full scale of five or more items would be the subject of a more detailed MOP development at some future time.

Two types of MOPS are presented (in separate sections)- behaviourally based subjective ratings by SMEs and objective performance data in terms of performance errors/accuracy and time to conduct certain tasks.

The measures are organised below according to mission phase, as outlined in the following summary table. Note that OP=Objective performance measures, BARS=Behaviourally Anchored Ratings Scales and PF=Pathfinder. The numbers in the BARS and OP columns refer to the specific measurement metrics that follow.

Measurement Concept: Planning and Briefing

Behaviour Domains	Behaviour Categories	Specific Behaviours	Measures		
			OP	BARS	PF
Detailed Planning	Tactics	Task understanding ROE understanding Route review/analysis Tactical effectiveness of plan Factors considered in plan COA considered Decision quality/timeliness Use of resources Creation of extra materials	1A 1A 1B. C	1A 1A 1A 1B 1C 1C 1D 1E 1F	
	Time Management	Time appreciation Time required Efficiency	2A--- 2E	2 2 2	
	CRM-Function Allocation	Clarity and assignment of team roles	3A 3B	3	
Planning Products		Quality of: Fuel plan TOLD Communications Mission data		 4A 4B 4C 4D	
Briefings	Communication	Detail Participation Comprehensiveness Time required	4A---- 4D	5 5 5 5	
	General effectiveness	Degree of instructor intervention		6	
Mental Model of mission	Knowledge Structures	Change in KS resulting from training			■

Measurement Concept: Mission Execution

BEHAVIOUR DOMAINS	BEHAVIOUR CATEGORIES	SPECIFIC BEHAVIOURS	MEASURES		
			OP	BARS	PF
Mission conduct	Plan compliance	Navigation accuracy Time control Communications required Compensation/adjustment Loss of separation incidents	5A 5B,C 5D 5E 5F	7A 7B	
	Communications	Speed/accuracy Terminology Necessity Information content	6A-E 6F,G 6H 6I	8	
	Aircraft handling/control	Piloting skills (e.g. accuracy in heading, speed, altitude, clearances)		9	
Situation awareness	System awareness	Checklist performance	7A-7D	10A	
		Sensors, status indicators	8A,B	10B	
	CRM	Crew awareness and communication	9A-E	11	
	Tactical awareness	Contact detection(speed,accy) Spatial awareness Co-ordination COA	10A-D 10E-H	12 12 12 12	
	Mission/goal awareness	Resumption of plan after engagement Detects changes in mission picture Changes/updates plan re changes in mission picture Communicates with crew	11A 11B 11C 11D	13 13 13 13	
	Flying skills	Engagement skills Weaponry skills Threat exposure Role discipline	12A-G 13A-F 14A-F 15A-C	14A 14B 14C 14D	
Goal Accomplishment	Achievement of objectives	Completeness, survival		15	
	General effectiveness	Degree of instructor intervention		16	
Mental Model	Knowledge structures	Team mental model/co-ordination/CRM			■

Measurement Concept: Debriefing

BEHAVIOUR DOMAINS	BEHAVIOUR CATEGORIES	SPECIFIC BEHAVIOURS	MEASURES		
			OP	BARS	PF
Briefings	Communication	Detail, Participation, Comprehensiveness, Time spent	12A	17	
	Mission accomplishment	Time spent Completeness	12B	18	
	Crew CRM	Time spent Completeness	12C	19	
	Crew Technical	Time spent Completeness	12D	20	
	Lessons Learned	Time spent Relevancy	12E	21	
	General effectiveness	Degree of instructor intervention	12F	22	

The following sections contain the detailed BARS followed by the objective performance measures.

Behaviourally Anchored Ratings Scales (BARS)

Mission Planning and preparation

MOPs are organised according to the following aspects of mission planning that have been found to best predict mission performance.

- Tactics
- Time management
- Function allocation and Crew Resource Management (CRM)
- Planning products
- Communication

The specific set of MOPs will need to be tailored for each particular circumstance. In particular, the measures were developed with the understanding that there is adequate time for planning and preparation. If, however, such time is scarce, as in actual operations with long mission planning cycles and the large manpower demands of a 24/7 modern air combat campaign, (Top Aces, 2003), then the list of proposed measures will need to be severely truncated to reflect actual combat planning opportunities and processes.

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1A. TACTICS: TASK AND ROE UNDERSTANDING, ROUTE REVIEW/ANALYSIS,				
Accepts mission tasking without review of important information	Accepts mission tasking with some review of key information	Good review of important information and understanding of the mission tasking	Carefully analyses the mission tasking and important information	Meticulously analyses the mission tasking and all related information
Look for: Thorough review of Air Tasking Order (ATO), Airspace Coordination Order (ACO), Intelligence update Thorough review of target area and routing Thorough review of ROE, FLIP, ASU, NOTAMS Threat situation/update Aware of significant terrain along planned routing Ingress/egress corridors (FOR DMT) ensure all non-site participants included where appropriate? **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1B. TACTICS: FACTORS CONSIDERED IN PLAN, TACTICAL EFFECTIVENESS OF THE PLAN, DEVELOPMENT OF MISSION PRODUCTS				
No consideration of key planning factors Very poor tactical plan selection No additional products added to mission execution plan	Some discussion of planning factors Poor tactical plan chosen using some of the available resources Minimal products added to mission execution plan	Identifies and analyses many key planning factors Good tactical plan chosen Incorporates some products into mission execution plan	Identifies and analyses all key planning factors Very sound tactical plan chosen using all resources Integrates many additional products into the mission execution plan Challenge/question plan assumptions	Provides a detailed analysis of all key planning factors Ideal and creative tactical plan selected using all available resources and inputs All products are incorporated into the mission execution plan

Look for:

Were key planning factors or inputs omitted?

Was there due consideration of all factors in tactical plan selection?

Was the plan selection discussed with other crewmembers and were valuable inputs considered?

Tactics considered for each mission phase?

Consider the big picture?

Detailed threat analysis?

Terrain masking

Ingress/egress corridor selection

Crew members planning responsibilities defined

Terminal operations tactics selection (??? Consider revising/deleting)

(FOR DMT) ensure all non-site participants included where appropriate?

Observations:

Rate the style of plan chosen: strongly conservative, conservative, neutral, aggressive, strongly aggressive.

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1C. TACTICS: COA AND CONTINGENCY PLAN CONSIDERS PERFORMANCE FACTORS, MALFUNCTIONS, WEATHER, ALTERNATE AIRFIELDS				
Accepts mission plan with no consideration of "what-if" contingencies	Some discussion of "what-ifs" No attempt made to incorporate contingency options into mission plan	Identifies some "what-if" contingencies Incorporates these into mission plan	Identifies many "what-if" contingencies Incorporates most into mission plan	Provides detailed consideration of "what-if" contingencies Integrates these into mission plan
Look for: Plans for what if's in the primary mission plan Plans for secondary / alternate missions and applicable what if's Plans for battle damage, missing members, late arrivals Plans departures / arrival contingencies (ex. go around) Weather contingencies / abort plan/ alternate airfields Alternate tactics considered for each mission phase? **Observations:** **Rate the style of contingency plan chosen: strongly conservative, conservative, neutral, aggressive, strongly aggressive.**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1D. TACTICS: DECISION QUALITY: QUALITY AND TIMELINESS OF DECISIONS				
No decisions made ahead of time	Made one decision ahead of time Little rationale	Made several decisions at the appropriate time Some rationale	Made most possible decisions at the appropriate time Identified other possible ones	Made all possible decisions at the optimum time Identified other possible ones
Look for: Decisions being made at the optimum time and before they are required Decisions are correct Involvement of other crew members and all available resources in decision making Communicated decisions in a clear and timely fashion **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1E. TACTICS: USE OF PLANNING MATERIALS: REFERENCE TO PLANNING ITEMS IN MISSION PLANNING KIT				
Little or no reference to available planning materials in constructing mission plan	Periodic reference to some materials in constructing mission plan	Frequent reference to some of the materials in constructing mission plan	Extensive reference to most of the available materials in constructing the mission plan	Extensive reference to all available materials in constructing the mission plan
Look for: Use of SOPs Use of pre-planned material / tools Use and adherence to current tactics manual / Airplane Operating Manual Use of applicable computer programs Check NOTAMS Ask for Flight Weather Briefing Check Airfield Suitability Update Check sketches for accuracy Request specific resources required for planning? Ferret out needed materials/resources? **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
1F. TACTICS: DEVELOPMENT OF PRODUCTS: CREATION OF EXTRA MATERIALS TO HELP MISSION UNDERSTANDING				
Creates no additional materials or products	Creates at least one additional product Minimal detail	Creates two additional products Some detail	Creates several additional products Considerable detail	Creates a number of additional products Extensive detail
Look for: Relevance and clarity of extra material (is it useful?) Target area photos, satellite imagery, FLIR pictures Route sketches, stick diagrams Comm card annotations Fire control area sketch, map annotations Makes extra copies of maps for other CMs **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
2. TIME MANAGEMENT: TIME APPRECIATION, EFFICIENCY IN TIME SPENT PLANNING TO ACCOMPLISH ALL REQUIRED PLANNING ACTIVITIES				
No time appreciation carried out Spends minimal time planning Much of the time is unproductive	Cursory time appreciation carried out Spends more than the minimal time planning but not much Considerable unproductive time	Time appreciation carried out Spends an adequate time planning Does not use all available time	Good time appreciation considering most inputs Spends considerable time planning Makes effort not to waste time during the planning session	Ideal time appreciation made with due consideration for non-standard timings and inputs Uses all available time planning Little or no wasted time during the planning session
Look for: Adequate time appreciation made with due consideration for non-standard inputs Time appreciation and plan communicated to other crew members Time appreciation monitored and revised during planning process Note how much time was actually spent planning Note how much time was wasted during planning Indicate which CMs were planning and which were not **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
3. Function allocation and CRM				
AC provides no direction for planning responsibilities Crew engages in planning with no discussion of roles	AC gives limited direction for planning responsibilities and expectations	AC gives good direction to all crew for planning responsibilities	AC provides clear direction to all crew for planning responsibilities and guidance on time to be spent in planning AC outlines expectations for planning products AC reviews crew products	AC provides very clear direction to all crew for planning responsibilities and guidance on time to be spent in planning AC monitors planning process of crew and provides timely and appropriate feedback AC reviews all crew products and all errors are rectified
Look for: Responsibilities are delegated by the AC iaw SOPs Responsibilities are delegated according to individual strengths Expectations of AC are clear Crew members planning activities are well supervised Planning products are well reviewed, errors are found, explained and corrected AC able to supervise crew members without sacrificing own responsibilities **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
4a. Quality of planning products: Fuel plan: level of detail, quality of plan, use of computer products				
Accepts fuel plan as given with no further observation	Checks fuel plan Makes no additional adjustments	Checks fuel plan Considers making some adjustments based on additional planning factors (airfield elevation, ACL)	Considers fuel plan against a range of options (e.g., engine loss, medical mission, Wx change)	Includes all mission critical and environmental variables in fuel plan
Look for: Considers primary plan and contingencies (what if no air-to-air refuelling, weather enroute, engine failure enroute)Compares fuel load against ACL (??? not familiar with term ACL), medevac pax **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
4B. QUALITY OF PLANNING PRODUCTS: TOLD: CONSIDERATION OF AIRFIELD INFO, OBSTACLES, PERFORMANCE FACTORS, NOTAMS, FUEL LOAD				
Computes TOLD with major errors and omissions	Computes TOLD with several errors and/or omissions	Computes TOLD with minimum errors Minimum mission requirements addressed	Computes TOLD with no errors Most mission variables addressed	Computes TOLD with no errors All mission variables addressed
Look for: Computes TOLD cards for other possible landing areas Is pilot helping FE identify TOLD? Is TOLD for future events based on future meteorlogical data, or planned using current data? Updates in weight, weather, runway surface conditions, barometric pressure and temperature considered? Is TOLD calculated for contingencies (backup aircraft configurations, fuel loads, weapons loads) **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
4C. QUALITY OF PLANNING PRODUCTS: COMM PLAN: LEVEL OF DETAIL, QUALITY OF PLAN, USE OF COMPUTER PRODUCTS				
Accepts comm plan as given with no further consideration	Checks comm plan Makes no additional adjustments	Checks comm plan Considers making some adjustments based on additional planning factors (IFR/VFR, high elevation)	Considers comm plan against a wide range of options (e.g., fire control area, emergency airfields)	Includes all mission critical and environmental variables in comm plan
Look for: Understands complex C2 structure Indicates who has OPCON at key points Considerations for alternate comm plan, secure requirements, monitoring of common / important frequencies, comm jamming, comm failure **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
5. COMMUNICATION: MISSION BRIEFINGS: DETAIL, PARTICIPATION, COMPREHENSIVENESS, OVERALL EFFECTIVENESS				
Little or no mission briefing. Little or no crew participation Time wasted on irrelevant communication Overall plan not explained	Minimal mission briefing and crew participation. Most elements covered. Provides little detail on objectives Overall plan not well explained	Brief contains some inconsistent mission details Few crew participate Overall plan explained with some detail	Detailed briefing Several crew participate Many elements covered with acceptable level of detail Overall plan well explained and understood	Highly detailed mission brief. All crew participate All mission elements covered with great detail Overall plan very well explained, communicated and understood by all crew members

Look for:

Does AC involve crewmembers and encourage participation?

() (Suggestion: this belongs in the Detailed planning phase)

Was key info omitted?

Were there parts of the brief requiring more detail?

Was the time available well used to cover most critical items of the mission and were they well prioritized?

Was it rushed?

Did it ramble (unfocussed)?

Was the actual tactical plan well explained?

Does crew appear to leave briefing with understanding of mission? - with confidence?

(For DMT) were there appropriate communications with remote mission participants?

Note duration of briefing:_____ (mins)

Observations:

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
6. OVERALL EVALUATION: DEGREE OF INSTRUCTOR INTERVENTION: DEGREE OF ASSISTANCE RENDERED TO ACCOMPLISH PLANNING EVENTS				
Extensive assistance is provided for the crew to successfully accomplish planning events	Substantial assistance is provided for the crew to successfully accomplish planning events	Limited assistance is provided for the crew to successfully accomplish planning events	Coaching (for technique only) is provided for the crew to successfully accomplish planning events	No assistance is required or provided Minimal errors made and corrected by the crew and/or crewmember when discovered
Look for: Interventions required to keep crew "on track" during planning session After delegation, instructor having to take on some planning tasks to assist weaker crewmembers Does the AC take charge of planning process and assume instructional role for other crew members? **Observations:**				

Additionally, the aircrew as a whole and as individuals should be rated for their overall planning behaviour according to the following scale:

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
OVERALL PLANNING BEHAVIOUR				
Observed planning behaviours were significantly below expectations.	This level of proficiency is less than desired for effective coordination during the mission. There is room for much improvement.	The demonstrated behaviour promotes and maintains coordination and mission operations effectiveness. This is the minimum standard level of proficiency that should be expected during this mission.	Observed planning behaviours are significantly above expectations.	Behaviours represent a high level of skill in the application of specific behaviours, and serves as a model for coordination, teamwork and highly efficient mission operations.
Note any exceptionally positive or negative behaviours observed:				

THIS PAGE INTENTIONALLY LEFT BLANK

Mission Execution

The evaluation of mission execution performance is broken down into the following elements:

- Plan compliance (includes navigational accuracy and timing control)
- Communications
- Aircraft handling and control
- Situation awareness: system, crew, tactical, mission/goal
- Achievement of objectives

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
7A. PLAN COMPLIANCE[5]: NAVIGATION ACCURACY: AWARENESS OF CURRENT LOCATION, ADHERENCE TO PLAN CONSIDERED				
Crew is lost/disoriented No adherence to planning routing Unable to meet objective requirements	Often deviates from / is unsure of routing Only able to meet objective requirements with difficulty	Generally able to follow the planned routing Several off-track deviations performed to meet objectives	Is able to adhere to planned mission routing Is aware of position with respect to objectives at all times	Is continually aware of position and anticipating routing ahead Able to make timely and appropriate adjustments to meet objective requirements
Look for: Print (HC) cross-country ground track from IOS at each waypoint (???) Note any crew discussions indicating confusion over present position Accuarate position (Ex.:Should stay within required (1.5mi) distance of each waypoint) Appropriate routing corrections adjustments made Crew adapts to the current situation Crew anticipation of routing ahead Perform adequate checks on nav data entered into flight computer **Observations:**				

[5] Note that this evaluation also covers factors identified under the heading "*Spatial, temporal and identity awareness*" in the main body of the report

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
7B. PLAN COMPLIANCE: TIME CONTROL: STAYS IN SYNCH WITH PLANNED TIME MILESTONES				
Misses many important time milestones Is continually "behind the mission" throughout	Misses several key time milestones	Hits all major time milestones Misses several minor ones	Hits all time milestones	Hits all time milestones and is continually "ahead of the mission" throughout
Look for: Note mission start time Note stations time Hit ETAs within appropriate time deviation. Ability to analyse and correct timing deviations **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
B. COMMUNICATIONS SYSTEMS USAGE: STAY ON CORRECT FREQUENCY, TALK TO PROPER AGENCY, CORRECT TERMINOLOGY				
Frequent problems in contacting proper agency or using correct frequency	Periodic problems in contacting proper agency or using correct frequency	No major problems in contacting proper agency or using correct frequency	Proper agencies contacted, correct frequencies, code words and call signs and used throughout	Clear, efficient communications executed throughout entire mission
Look for: Correct frequency, correct agency used/contacted throughout Get all clearances when needed Use of appropriate call signs? Use of appropriate code words? Correct terminology? Concise phrasing? Transmit only when situ required? Responds appropriately and promptly when contacted by external agency/formation member(s) Crew question communication when not in agreement with what is being observed Properly monitors all appropriate frequencies No broadcast or specific information missed **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
9. AIRCRAFT HANDLING AND CONTROL: AIRSPEED, ALTITUDE, HEADING				
Excessive flight path deviations which are not corrected in a timely manner	Some flight path deviations which are not all corrected in a timely manner	Few flight path minor deviations which are corrected in a timely manner	Minimal flight path deviations which are corrected in a timely manner	Minimal flight path deviations which are anticipated and corrected in a timely manner

Look for:

Maintains adequate terrain clearance.

Maintains safe separation from other aircraft.

Maintains correct flight parameters (ground speed, altitude, heading) throughout the mission.

Speed/accuracy of corrections to observed errors in aircraft flight parameters.

Speed/accuracy of changing flight parameters when requested by external agencies (ATC, AWACS, Flight Lead).

Anticipation of required changes

Optimum aircraft performance attained and maintained

Degree to which over-corrections to observed errors are made

Observations:

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
10A. SYSTEM AWARENESS: CHECKLIST ACCOMPLISHMENT: CHECKLISTS ACCOMPLISHED IN A TIMELY, ACCURATE MANNER				
Fails to complete many Emergency and/or normal procedures/ checklists	Most Emergency and/or normal procedures/checklists complete Not timely or misses items	All required Emergency and/or normal procedures/ checklists complete	All required Emergency and/or normal procedures/ checklists are completed in timely manner Covers all items	All Emergency and/or normal procedures/ checklists completed in timely manner or early Efficiently covers all items
Look for: Ability to perform normal procedures items by memory without mistakes Complete before start checklist, before taxi checklist, taxi checklist, before takeoff checklist, arming point checklist, line up checklist on time Completes all tactical checklists properly and at the appropriate time (fence-in check, Go/No Go checklist, air-to-ground checklist, fence-out checklist) Complete descent checklist, before landing checklist, after landing checklist, and de-arming checklist on time Completes all necessary emergency checklists (if required) in a timely, safe, and efficient manner Completes all emergency memory / recall items timely and accurately **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
10b. System awareness: sensors, status indicators				
Fails to recognise critical changes in sensors or status indicators	Slow to recognise critical changes in sensors or status indicators	Recognises and responds to critical changes in status or sensors	Integrates and correlates information between sensors	Uses available information quickly and efficiently according to an expert strategy
Look for: Awareness of overall aircraft status and indications including all operational, degraded, and unserviceable items (fuel, flight controls, avionics, aircraft subsystems, weapons) Awareness of sensor status and information (radar, FLIR, RWR, data link, jammers, IFF interrogator, etc). Awareness of what impact specific changes to aircraft or sensor status will have on overall mission. Amount of time required to notice changes to aircraft system/sensor status. Ability to correlate information between different sensors / indicators **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
11. Resource and crew awareness				
Fails to detect changes in crew status or mission resources	Misses some changes in crew status or mission resources	Detects critical changes in crew status or mission resources	Rapidly detects critical changes in crew status or mission resources Responds to crew communications re status Plans/compensates for changes in crew status or mission resources	Proactively scans for info re changes in crew status or mission resources Clearly sets/reminds crew roles
Look for: Ability to detect formation members divergence from mission plan Out of formation not detected Awareness of position and actions of other formation members throughout the mission Awareness of status of critical mission resources and actions taken when mission resources change Communications re crew status missed (For Mission Commander/OC) no directions to crew out of formation or not adhering to the plan Accurate and timely directions given to crew members / mission resources with changes in tactical situation **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
12. TACTICAL AWARENESS: CONTACT DETECTION, SPATIAL AWARENESS, CO-ORDINATION, MONITORS FORMATION, APPLIES OTHER COA				
Fails to detect contact in timely manner Loses sight of friendly/enemy entities Engagement quality drops during comms Cannot asses and fly at same time	Frequently out of position Employs incorrect tactics Does not co-ordinate with formation	Detects at appropriate range Recognises enemy tactics Adopts correct tactics	Detects enemy at earliest opportunity Monitors formation Adopts appropriate tactics to changing tactical situation	Flies optimum engagement profile Co-ordinates effectively with formation Anticipates enemy
Look for: Ability to detect contacts with sensors Awareness of position relative the mission routing, enemy formations, friendly assets Ability to understand local tactical environment and choose effective COAs Ability to monitor and direct other formation members COAs Awareness of enemy formations actions, and anticipation of outcome Not setting/maintaining roles Not flying best 1v1 as engaged fighter Free fighter losing sight **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
13. MISSION AND GOAL AWARENESS: RE-ESTABLISHES MISSION GOALS, DETECTS AND RESPONDS TO CHANGES IN MISSION PICTURE				
Fails to detect changes in mission situation following engagement Poor resumption of nav plan/join of formation following engagement	Recognises change in mission picture Unsure of appropriate COA in response to change Unsure of mission goals at various points in mission	Recognises changes in mission situation and adjusts appropriately	Integrates information to quickly recognise changes in mission picture Rapidly updates plan and communicates changed picture and plan	Anticipates changes in mission picture Has contingency mission goals
Look for: Maintaining a broad scan of info sources (i.e. radio, hud, outside, etc.). Ability to comprehend current changes in the tactical environment. Ability to anticipate the effect of current changes in the tactical environment on future events in the mission. Ability to develop contingency COAs in response to the potential impact of current mission events Appropriate assessment of tactical situation and use of defensive, offensive and neutral manoeuvring tactics **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
14A. ENGAGEMENT SKILLS: AIRCRAFT HANDLING, ENERGY MANAGEMENT, GAINED/MAINTAINED OFFENSIVE ADVANTAGE				
Constant and large deviations from ideal energy state Quickly lost clearly offensive position Was ineffective to counter bandits offensive	Many deviations from ideal energy state Maintained position but was unable to fly to a weapon engagement zone or to leave the engagement	Good energy management and A/C manoeuvering Maintained offensive position and employed weapons Countered bandits offensive	Very sound energy management and aircraft manoeuvering Quickly capitalized on offensive position Effectively countered bandits offensive to a neutral position	Ideal Aircraft energy management and manoeuvering Expeditiously capitalized on offensive position Quickly reversed from a defensive to an offensive position
Look for: Aircraft airspeed, g, and angle of attack Reactions to bandits manoeuvering and energy state Appropriate game plan used Recognition of turning room Lift vector placement Divergence for ideal energy state Awareness of altitude Quick exit after engagement or when opportunity presented **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
148. WEAPONRY SKILLS: RECOGNITION OF WEAPONS EMPLOYMENT OPPORTUNITIES, SATISFIED ROE, VALIDITY OF SHOTS AT TRIGGER SQUEEZE				
Did not recognize entry into WEZs Failed to employ weapons according to the briefed plan Most shots did not meet all shoot criteria at trigger / release	Did not recognized some weapons employment opportunities Generally did not follow weapon employment plan Some shots did not meet weapons release criteria	Recognized entry into most WEZs Generally employed weapons according to briefed criteria Met all weapon shoot criteria at trigger / release	Recognized all WEZs Employed weapons according to briefed criteria Met all weapon shoot criteria at trigger / release Appropriately assessed requirement for follow on shots	Anticipated and recognized all WEZs opportunities Always employed optimum weapon according to briefed first / ideal opportunity All criteria met for all weapon releases Always anticipated and recognized requirement for follow shots Recognized all degraded / Lower PK situations
Look for: WEZ (Weapon Engagement Zone) WEZ entries and anticipation Missed weapons opportunities Proper weapon selected and employed Weapon Pk at trigger / release (% of RNE, Range No Escape) All release criteria met at trigger Recognition of follow shot requirement **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
14c. Threat Reactions: recognition of exposure to threats, effectiveness of threat reactions, knowledge of threat performance/capabilities				
Failed to avoid or recognized most threats Did not perform appropriate counter manoeuvers to threats Did not react appropriately when enemy employed weapons	Recognized most threats When threat recognized, did not always perform appropriate counter manoeuvers Did not react to all enemy weapons opportunities	Avoided some threats Recognized all threats Performed appropriate counter manoeuvers Reacted to enemy weapons opportunities	Avoided most threats Quickly recognized all threats Effectively performed appropriate counter manoeuvers Denied enemy of most weapons opportunities or quickly reacted accordingly	Avoided all non-necessary threats Anticipated all potential threats and proactively performed ideal counter manoeuvers Denied enemy of weapons opportunities
Look for: Anticipation and recognition of threats Pre emptive counter manoeuvrers Appropriate counter manoeuvers performed when approaching threats Effectiveness of Counter manoeuvers Enemy weapons opportunities Ability to assess follow on threats when reacting Employment of appropriate counter measures (Chaff, Flares, Jammers, IRCCM) **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
14D. ROLE DISCIPLINE: ABILITY TO FULFILL ASSIGNED ROLE WITHIN THE MISSION, ABILITY TO FOLLOW THE BRIEFED PLAN, ABILITY TO SUPPORT OTHER CREW/FORMATION MEMBERS DURING MISSION				
Unable to perform or was ineffective at most individual responsibilities Unable to support others or negative contribution Mostly contributed negatively by assuming wrong responsibilities Produced chaos in the formation	Missed some critical individual responsibilities Partially supported others when required Sometimes contributed negatively by assuming unnecessary responsibilities	Performed all individual responsibilities Supported others most of the time Was able to assume other role responsibilities some of the time	Performed all individual responsibilities with high standard Timely supported others when required Effectively assumed other role responsibilities when required	Ideally and timely performed all individual responsibilities Anticipated degrading situations or opportunities and immediately supported others when required Instantly and ideally assumed other role responsibilities when required
Look for: Performance of individual responsibilities Assertiveness in role Respecting other roles responsibilities Timely response and support Ability to perform other roles, additional duties Overall contribution to mission success **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
15: ACHIEVEMENT OF PRIMARY MISSION OBJECTIVES				
Primary mission objective(s) not accomplished Did not survive the mission	Some mission objective(s) not accomplished Survivability jeopardized on several occasions	Mission objective(s) satisfactorily accomplished Minimal threats to survivability	Mission objectives easily accomplished Did not place survivability of formation at risk	All mission objectives accomplished in an optimum, safe, efficient manner Ensured survivability of own and other mission assets
Look for: Clear understanding of mission objectives Awareness of what are the critical tasks to perform in order to accomplish mission objectives Awareness of how mission events might affect accomplishment of mission objectives Unnecessary exposure to enemy threats, terrain, and close proximity of other aircraft that jeopardizes the survivability of the formation Ability to perform self analysis of progress vs. objectives and adjust COA's accordingly **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
16: DEGREE OF INSTRUCTOR INFLUENCE: DEGREE OF ASSISTANCE RENDERED TO ACCOMPLISH MISSION EVENTS				
Extensive assistance is provided for the crew to successfully accomplish mission events	Substantial assistance is provided for the crew to successfully accomplish mission events	Limited assistance is provided for the crew to successfully accomplish mission events	Coaching (for technique only) is provided for the crew to successfully accomplish mission events	No assistance is provided Errors are corrected by the crew and/or crewmember when discovered
Look for: Instructor actions required to keep crew "on track" during mission Ability to recognize errors, quickly rectify the situation Ability to teach others **Observations:**				

Additionally, the aircrew as a whole and as individuals should be rated for their overall mission execution behaviour according to the following scale:

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
OVERALL MISSION EXECUTION BEHAVIOUR				
Observed mission execution behaviours were significantly below expectations.	This level of proficiency is less than desired for effective coordination during the mission. There is room for much improvement.	The demonstrated behaviour promotes and maintains coordination and mission operations effectiveness. This is the minimum standard level of proficiency that should be expected during this mission.	Observed behaviours are significantly above expectations.	Behaviours represent a high level of skill in the application of specific behaviours, and serves as a model for coordination, teamwork and highly efficient mission operations.
Note any exceptionally positive or negative behaviours observed:				

Mission Debrief

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
17. Overall quality and communication				
No debrief provided or only cursory mission debriefing with little or no crewmember participation	Debrief provided Critical issues left unresolved Some crewmember participation in debrief	All debrief elements covered with acceptable level of detail for mission objectives and crewmember participation	Detailed debrief with all mission elements covered Substantial participation by most crewmembers	Highly detailed debrief covering all mission elements Full crew participation
Look for: Record how long the debriefing lasted? Was the debrief length appropriate to the complexity of the mission and the potential lessons learned (Note: this may be approximate since they may start debriefing on their way back from the sim) Note whether it was instructor-led or led by the crew Audio-video/playback tools: did crew take advantage of capabilities to review their own technical/CRM performance? Does the debrief include discussion of all relevant events that influenced the outcome/success of mission? Does the debrief follow a logical review of mission events? Was the reconstruction of the mission events accurate? Did the AC involve crew when required? (Interaction of the crew during reconstruction and debrief) **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
18. MISSION OUTCOMES: FOCUS ON ACCOMPLISHMENT OF MISSION OBJECTIVES				
Fails to accomplish a debrief of mission outcomes Does not identify any critical mission deficiencies	Accomplishes a minimal debrief of mission outcomes Identifies few mission critical deficiencies	Accomplishes an acceptable debrief of mission outcomes Identifies only critical mission deficiencies	Accomplishes a debrief of mission outcomes Identifies some mission deficiencies and some non-critical issues	Accomplishes a thorough and accurate debrief of mission outcomes Identifies all critical and some non-critical deficiencies
Look for: Does debrief centre on determining whether mission objectives were achieved? Is a quantitative assessment conducted to assess overall mission objective success? Are valid reasons for failure to achieve specific mission objectives provided? Are the lessons learned recognized and well explained? (How to do it better) Focus on improving performance Consideration of alternative approaches to mission **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
19. Crew CRM performance: focus on CRM performance of the crew				
Fails to accomplish a debrief of each CMs CRM performance Does not identify any critical performance deficiencies	Accomplishes a minimal debrief of each CMs CRM performance Identifies few critical performance deficiencies	Accomplishes a debrief of each CMs CRM performance Identifies only critical performance deficiencies	Accomplishes a debrief of each CMs CRM performance Identifies critical & some non-critical performance deficiencies	Accomplishes a thorough and accurate debrief of each CMs CRM performance Identifies all critical & non-critical performance deficiencies
Look for: Identify and expand on areas of the mission where CRM was strong/weak. Recognized lessons learned and proposed and discussed different methods to improve CRM. Are critical teamwork incidents covered? Is 'no blame' culture maintained? **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
20. Crewmember technical performance: focus on technical performance of crewmembers				
Fails to accomplish a debrief of CM's technical performance Does not identify any critical performance deficiencies	Accomplishes a minimal debrief of CM's technical performance Identifies few critical performance deficiencies	Accomplishes a debrief of each CM's technical performance Identifies only critical performance deficiencies	Accomplishes a debrief of each CM's technical performance Identifies critical and some non-critical performance deficiencies	Accomplishes a thorough and accurate debrief of each CM's technical performance Identifies all critical and non-critical performance deficiencies
Look for: Is there a distinction between crew technical performance and crewmember performance? Are technical errors recognized and discussed throughout the debrief? Is the debriefing of technical performance accurate yet efficient? Are ways to improve provided to minimize errors in the future? **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
21. LESSONS LEARNED: FOCUS ON LESSONS LEARNED DURING THE MISSION				
Did not identify any lessons learned, or lessons learned were irrelevant/inaccurate	Attempted to identify valid lessons learned, but they were largely irrelevant or contained some degree of inaccuracy	Identified some valid lessons learned from the overall mission.	Identified valid lessons learned for the overall mission, as well as specific lessons learned related to key events during the mission	Identified all key lessons learned from the mission. Clearly explained what caused the lesson learned to occur and how to improve performance in the future
Look for: Important lessons learned missed related to overall mission conduct Important lessons learned missed specific to significant mission phases and events Was well explained/understood why the lesson learned occurred, and how to improve performance in the future? Were lessons learned identified that were inaccurate or irrelevant? **Observations:**				

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
22. INSTRUCTOR INTERVENTION				
Extensive assistance is required for crew to successfully accomplish debrief events	Substantial assistance is required for crew to successfully accomplish debrief events	Limited assistance is required for crew to successfully accomplish debrief events	Coaching (for technique only) is required for crew to successfully accomplish debrief events	No assistance is required Debrief errors are corrected by the crew and/or crewmember when discovered
Look for: Do the instructors have to "goad" the crew into talking about the mission? **Observations:**				

Additionally, the aircrew as a whole and as individuals should be rated for their overall debriefing behaviour according to the following scale:

1. Poor	2. Marginal	3. Standard	4. Very Good	5.Exceptional
OVERALL DEBRIEFING BEHAVIOUR				
Observed debriefing behaviours were significantly below expectations.	This level of proficiency is less than desired for effective coordination during the mission. There is room for much improvement.	The demonstrated behaviour promotes and maintains coordination and mission operations effectiveness. This is the minimum standard level of proficiency that should be expected during this mission.	Observed debriefing behaviours are significantly above expectations.	Behaviours represent a high level of skill in the application of specific behaviours, and serves as a model for coordination, teamwork and highly efficient mission operations.
Note any exceptionally positive or negative behaviours observed:				

This approach to overall aircrew assessment leads to the following instrument. The SME assessing the training exercise should make an overall assessment for the whole crew for each mission segment based on the relevant 5-point scale (see sections on individual mission segments). The SME should circle, or put a check-mark, or an 'X' in the corresponding box.

The SME assessing the training exercise should make an assessment for each individual crewmember for each mission segment based on the relevant 5-point scale (see sections on individual mission segments). The SME should circle, or put a check-mark, or an 'X' in the corresponding box.

	Planning					Execution					Debriefing				
Entire Crew	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Flight Lead	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Pilot 1	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Pilot 2	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Pilot 3	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5

Objective Performance Data

Mission Planning

Tactics

1A. Time spent reviewing tasking order, ROE, intelligence
1B. Number of COA considered
1C. Number of contingency plans

Time Management

2A. Total time spent in planning
2B. Time left at end of briefing for questions
2C. Time left at end of briefing before mission execution
2D. Time spent by OC in briefing at start of planning
2E. Time spent by OC in briefing to review final plan

CRM-Function Allocation

3A. Proportion of team actively engaged in planning
3B. Flow of information. (evaluate directionality, crew involvement)

Briefing/Communication

4A. Objectives tabulated
4B. Total time spent in briefing
4C. Number of open vs. closed questions
4D. Number of collaborative/working questions vs. the number of confirmation questions

Mission Execution

Plan compliance

5A. Number of spatial deviations from plan that exceed criteria.
5B. Number of time deviations from plan that exceed criteria.
5C. Number of crosscheck times that exceed criteria.
5D. Number of communications to crew concerning plan compliance.
5E. Proportion of deviations from plan correctly compensated for by adjustments in flight profile.
5F. Number of separation deviations (Ground, Altitude, Other aircraft)

Communications

Speed/accuracy

6A. Number of delayed communications
6B. Number of requests for correction
6C. Number of requests for repetition
6D. Number of failures to acknowledge/confirm
6E Number of missed communications
6F. Number of departures from standard terminology
6G. Number of uses of non-standard terminology
6H. Number of unwarranted communications
6I. Number of errors in information content

Situation awareness: system

7A. Time to complete checklist
7B. Proportion of items omitted from checklist
7C. Number of failures to complete checklist at required time
7D. Number of wrong checklist used or failure to recognized problem
8A. Proportion of critical status changes undetected (any critical sensor, status indicator)
8B. Mean time to respond to critical status change

Situation awareness: resource/crew

9A. Proportion of time crew out position not detected
9B. Number of omissions in failing to communicate status changes to crew.
9C. (For OC) Proportion of requests from crew unanswered.
9D. (For crew) Number of failures to report critical changes in status.
9E. Number of times crew lost sight of formation

Situation awareness: tactical

10A. Total time not conducting assigned sensor search pattern
10B. Number of failures to detect contact
10C. Response time from contact first appearance to detection
10D. Number of times lost sight of contact
10E. Number of errors in positioning while engaging contact
10F. Number of missed opportunities for releasing weapons against contact

10G. Number of times within lethal envelope of contact
10H. Number of errors in co-ordinating with formation
10I. Number of errors in maintaining role/position

Mission Goal Awareness

11A. Number of errors in resuming plan after engagement
11B. Errors in detecting changes to mission picture
11C. Errors in updating plan due to changes in mission picture
11D. Number of communications to crew to indicate changes in mission picture

DEBRIEFING

12A. Total time spent in debriefing
12B Proportion of time spent in reviewing mission goals and accomplishment
12C. Proportion of time spent in reviewing CRM performance of crew
12D. Proportion of time spent in reviewing individual crew technical performance
12E Proportion of time spent in reviewing lessons learned
12F Number of interventions by instructor required to ensure adequate performance

Comprehensiveness

Using the mission sequence framework outlined in section 5.3 (STTO, TRP, ingress, attack, egress) it is possible to map the TTWODLDE mnemonic (Terrain, Target, Weather, Ordnance, Defences, Lateral Support, Delivery, Egress) onto each stage to ensure that the mission is comprehensively debriefed. The SME determines whether each topic is addressed for each mission stage and puts a check-mark in the appropriate cell, using a matrix as follows:

	Terrain	Target	Weather	Ordnance	Defences	Lateral Support	Delivery	Egress
STTO								
TRP								
Ingress								
Attack								
Egress								

The highest total possible for comprehensiveness is 40. Aircrew scores for the comprehensiveness of their debriefing are therefore their score out of 40. This score can be combined with the BARS for 'Overall Debrief' in order to record not only whether all topics were considered, but also how well they were considered.

THIS PAGE INTENTIONALLY LEFT BLANK

Annex B: Proposed methodology/trial plan for use of Pathfinder in evaluating aircrew training

In this section we outline the major steps and tasks to be done in order to conduct a trial with the Canadian Air Force using the Pathfinder methodology to assess changes in knowledge structures that occur as a result of formal training. The plan is broken down into two major areas: the specific steps that need to be performed and the logistical requirements necessary to support the trial.

1. Domain selection

Within the context of Canadian Airforce Training, there are presumably many courses that focus on different areas of skill development and knowledge acquisition. If the primary interest of the trial is on changes in knowledge structures as opposed to skills or procedures learning, then the focus needs to be on a course in which the learning of new concepts is a core component. In a DMT context, the focus could be on a course that prepares trained pilots for combat and tactical environments in a multi-national co-operative theatre.

Logistics

Review potential courses with SA and Airforce SMEs to generate a preliminary selection of candidate courses.

Determine from Air Force schedule timing of courses and potential access to pilot trainee populations to act as trial participants.

Obtain necessary Air Force buy in and support for conducting trial and specific permission for access to selected participant group

2. Concept selection

The process of concept selection requires that we narrow down within the broad training goals those concepts of primary interest that can be managed within the constraints of the pairwise comparison process that is central to the Pathfinder methodology. This should be a collaborative process between the SA, Air Force SMEs and the contractor. The goal will be to come up with a list of between 30-35 critical concepts. The CF-18 Training Task list and other training curricula and examinations should be useful data sources for assisting this process.

Logistics

Contractor identifies and recruits Air Force SMEs (? Top Aces)

Contractor or SA acquire relevant training material from Canadian Air Force.

Contractor develops and distributes methodology for concept selection

Contractor convenes meeting with SA, SMEs to select concepts using methodology circulated

3. Concept refinement

Ideally, the process of concept refinement should be performed by a different set of SME's than those that generated the concept list. However, logistically this may be difficult to achieve if the pool of SMEs is small or otherwise unavailable due to time pressures. Two approaches are possible given the time and resources available. The least preferred method but probable most efficient would be to have the SMEs who constructed the list to each perform and independent rank ordering of the concepts. The data would then be averaged and the top 25 would be selected. A second approach that would be more time consuming would be to have the group do pairwise ratings of the concepts and then use Pathfinder analysis to exclude the less relevant items. The preferred approach would be to have a set of experienced pilots or instructors, with current domain knowledge perform the Pathfinder sort. The ensuing selection of concepts would therefore have high validity, and moreover the PFNETs derived from the procedure could be used as a basis for assessing changes in knowledge structures of the test participants.

Logistics

Assuming the group from the concept generation stage is available, the meeting would continue to allow either rankings or pairwise discriminations to be conducted. Contractor develops necessary protocols for this.

Assuming access to a pool of Air Force trainers, the following steps would be required

Contractor and SA negotiate with Air Force for access to required personnel, including details of time and place.

Contractor develops test protocol.

Contractor travels to test site and administers protocol.

Contractor analyses data (whichever method) and reports outcome to SA and recommends final list of concepts for the main trial.

Contractor finalises list of concepts with SA.

4. Pre-training administration of pairwise procedure

The contractor travels to test site and administers the initial pairwise sorting of the concepts by the target test participants.

Logistics

Contractor and SA negotiate with Air Force for specific access to required test participants

Contractor develops Ethical Review for subject participation and submits to SA for approval

Contractor reviews physical facilities available at site for test administration

Contractor and SA arrange for access to specific facilities

Contractor develops specific test protocol

Contractor acquires necessary software to analyse Pathfinder data

Contractor travels to test site to administer protocol

Contractor and SA negotiate with Air Force to obtain access to participant background and course performance data

Contractor briefs Air Force personnel and test participants

Contractor administers protocol

Contractor analyses data

Depending on the length of the course, it may be more practical and cost effective for the contractor to remain on site for the post training administration. During this time some preliminary data analysis will be conducted.

5. Post-training administration of pairwise procedure

The most critical component of this is the timing. USAF experience shows that it should be not left as the last thing to be done on the last course day by the test participants, particularly if this is a Friday afternoon.

Logistics

Contractor and SA negotiate with Air Force for optimum timing of this task

Contractor administers protocol

Contractor participates in wash-up or debrief and returns to base

6. Statistical analysis and interpretation of data

The contractor develops a statistical analysis plan and submits to the SA for approval and comments. The contractor then analyses and interprets the data in accordance with the plan, recruiting additional, specialised SME assistance wherever required. Pathfinder analysis and other techniques outlined in the statistical plan will be used to assess the change in knowledge structures from the start to end of the course and to compare the PFNETs of the trial participants with those of experts.

Logistics

Contractor develops protocol for data security and reviews with SA

Contractor responsible for data security

Contractor recruits and manages specialised SME assistance for data analysis

7. Reporting

Contractor delivers report to SA outlining the major findings of the study, limitations, generalisability and need for follow-up. The contractor also submits a plan for such work. Contractor provides an oral briefing to SA and Air Force.

8. Detailed outline of resources required

A detailed cost proposal to conduct the trial has been delivered separately to the Scientific Authority in

				LABOUR		
		Project Director	or Consultant	Consultant	SME	Clerical
Work Item #1: Project Management						
	Minimum necessary to achieve objectives of SOW					
1.01	Project Start-up Meeting (at DRDC Toronto)	4	4			
1.02	Track cost and schedule of contract work items	2	4			
1.03	Attend progress review meetings (Assume 4 x 0.5hr phone)	2	2			
1.04	Document progress review meetings	2	4			
1.05	Project admin	4	4			
Work Item #2: Domain Selection						
2.01	Review options	8	4			
2.02	Review schedules and access	8	4		2	
2.03	Liaise AF	2	8			
Work Item #3: Concept Selection						
3.01	Identify and liaise with SMEs	4			4	
3.02	Acquire and review relevant training material	16	16		8	2
3.03	Develop and distribute methodology for concept selection	8	16	16		2
3.04	Select concepts in conjunction with SA and AF SMEs (Assume via phone, e-mail)	8	16	4	8	
Work Item #4: Refine Concepts						
4.01	Conduct initial refinement with AF SMEs	8	16		8	
4.02	Liaise with AF to identify SMEs and availability	2	8			
4.03	Develop test protocol	8	16	16		
4.04	Travel to AF site and admin of protocol (Assume Cold Lake, 2 nights, 3 people)	24	24	24		
4.05	Analyse data	8	8	24	2	4
4.06	Reports and recommends final list of concepts	8	16	4		2
4.07	Final selection of concepts with SA	4	4			
Work Item #5: Pre-Training Admin of Pairwise Procedure						
5.01	Negotiate with AF for sepcific access to required test participants	2	8		2	
5.02	Develops Ethical Review for subject participation and submits to SA for approval	4	4	16		
5.03	Review physical facilities available at site for test administration (Assume trip to Cold Lake, 1 person, 1 night)	4	16			
5.04	Arrange for access to specific facilities		4			
5.05	Develop specific test protocol	4	8	24		
5.06	Travel to site and administer protocol/handle logistics (Assume Cold Lake, 2 nights, 3 people)	24	24	24		
5.07	Preliminary analysis of data	8	8	24	2	4
Work Item #6: Post-Training Admin of Pairwise Procedure						
6.01	Travel to site and administer protocol/handle logistics/washup (Assume Cold Lake, 2 night, 3 people)	24	24	24		
Work Item #7: Statistical Analysis and Interpretation of Data		16	40	40	4	8
Work Item #8: Report		24	40	40		4
Work Item #9: Presentation						
9.01	Prepare and present data (Assume presentation at DRDC Toronto, half day incl travel)	8	8	4		2
	Total Hours	**248**	**358**	**284**	**40**	**28**

electronic format.

DOCUMENT CONTROL DATA SHEET

1a. PERFORMING AGENCY
Humansystems Inc., Guelph, ON. N1H 3N4

2. SECURITY CLASSIFICATION
UNCLASSIFIED
Unlimited distribution -

1b. PUBLISHING AGENCY
DRDC Toronto

3. TITLE
(U) Development of Generic Aircrew Measures of Performance for Distributed Mission Training

4. AUTHORS
Michael L. Matthews Tabbeus M. Lamoureux

5. DATE OF PUBLICATION
April 1 , 2003

6. NO. OF PAGES
140

7. DESCRIPTIVE NOTES

8. SPONSORING/MONITORING/CONTRACTING/TASKING AGENCY
Sponsoring Agency:
Monitoring Agency:
Contracting Agency : DRDC Toronto
Tasking Agency:

9. ORIGINATORS DOCUMENT NO.
Contract Report CR 2003-060

10. CONTRACT GRANT AND/OR PROJECT NO.
W7711-007694/001/TOR

11. OTHER DOCUMENT NOS.

12. DOCUMENT RELEASABILITY
Unlimited distribution

13. DOCUMENT ANNOUNCEMENT
Unlimited announcement

14. ABSTRACT

(U) Advances in technology have made simulation and, latterly, distributed mission simulation valuable additions to the training of aircrew. Simulation is widely accepted by the aviation community and much research exists to show the benefits and most profitable applications of simulation. Distributed mission training represents an enhancement to simulation although at this point it is unproven exactly what training objectives should be associated with it and what additional benefits will accrue when compared to traditional simulation or flying training. This project developed generic measures of performance for application to distributed mission training exercises. The application of these measures of performance will allow training organisations to make valid statements about the benefits of distributed mission training and informed decisions to be made regarding which training objectives to address through the use of distributed mission training.
Humansystems Incorporated® were tasked with reviewing literature provided by DRDC Toronto in order to identify potential measures of performance. In particular, the Scientific Authority was interested in measures of mission planning, mission execution, mission debriefing, situation awareness and the change in aircrew knowledge structures, as relevant to distributed mission training. A measurement model has been developed that includes a conceptual outline of a CF-18 mission, a behavioural hierarchy composed of domains, categories and specific behaviours, and a range of associated rating scales and objective measures. Additionally, a trial plan for the application of one particular measure (Pathfinder – description and measurement of knowledge structures) has been developed.

(U) Grâce aux avancées technologiques, la simulation et, par la suite, la simulation de mission à distance, sont devenues de précieux ajouts à l'instruction de l'équipage. La simulation est largement répandue dans le monde de l'aviation, et bien des recherches démontrent les avantages et les applications les plus profitables de la simulation. L'instruction de mission à distance représente une amélioration en matière de simulation, même si l'on ne connaît pas encore exactement les objectifs d'instruction qu'il faudrait y associer et les avantages supplémentaires qui en découleront par rapport à la simulation traditionnelle ou à l'entraînement au vol. Les responsables de ce projet ont élaboré des mesures du rendement génériques applicables aux exercices d'instruction de mission à distance. L'application de ces mesures du rendement permettra aux organismes de formation de faire des recommandations pertinentes sur les avantages de l'instruction de mission à distance et de prendre des décisions éclairées sur les objectifs d'instruction à viser dans le cadre de l'instruction de mission à distance.
Humansystems IncorporatedMD a été chargée d'examiner la documentation fournie par RDDC Toronto en vue de déterminer des mesures du rendement potentielles. En particulier, le responsable des questions scientifiques s'intéressait aux mesures relatives à la planification de mission, l'exécution de mission, le debriefing de mission, la connaissance de la situation, ainsi que les modifications au niveau des structures de connaissances de l'équipage, en ce qui concerne l'instruction de mission à distance. On a élaboré un modèle de mesure incluant un aperçu conceptuel d'une mission de CF-18, une hiérarchie du comportement composée de domaines, de catégories et de comportements particuliers, ainsi qu'une gamme d'échelles de notation et de mesures objectives connexes. On a également élaboré un plan d'essai pour l'application d'une mesure particulière (Pathfinder – description et mesure des structures de connaissances).

15. KEYWORDS, DESCRIPTORS or IDENTIFIERS

(U) training; distributed mission training; simulation